RESEARCH CENTER
Rennes - Bretagne-Atlantique

FIELD

Activity Report 2019

Section New Results

Edition: 2020-03-21

CAIRN Project-Team

6. New Results

6.1. Reconfigurable Architecture and Hardware Accelerator Design

6.1.1. Algorithmic Fault Tolerance for Timing Speculative Hardware

Participants: Thibaut Marty, Tomofumi Yuki, Steven Derrien.

We have been working on timing speculation, also known as overclocking, to increase the computational throughput of accelerators. However, aggressive overclocking introduces timing errors, which may corrupt the outputs to unacceptable levels. It is extremely challenging to ensure that no timing errors occur, since the probability of such errors happening depends on many factors including the temperature and process variation. Thus, aggressive timing speculation must be coupled with a mechanism to verify that the outputs are correctly computed. Our previous result demonstrated that the use of inexpensive checks based on algebraic properties of the computation can drastically reduce the cost of verifying that overclocking did not produce incorrect outputs. This has allowed the accelerator to significantly boost its throughput with little area overhead.

One weakness coming from the use of algebraic properties is that the inexpensive check is not strictly compatible with floating-point arithmetic that is not associative. This was not an issue with our previous work that targeted convolutional neural networks, which typically use fixed-point (integer) arithmetic. Our on-going work aims to extend our approach to floating-point arithmetic by using extended precision to store intermediate results, known as Kulisch accumulators. At first glance, use of extended precision that covers the full exponent range of floating-point may look costly. However, the design space of FPGAs is complex with many different trade-offs, making the optimal design highly context dependent. Our preliminary results indicate that the use of extended precision may not be any more costly than implementing the computation in floating point.

6.1.2. Adaptive Dynamic Compilation for Low-Power Embedded Systems

Participants: Steven Derrien, Simon Rokicki.

Previous works on Hybrid-DBT have demonstrated that using Dynamic Binary Translation, combined with low-power in-order architecture, enables an energy-efficient execution of compute-intensive kernels. In [33], we address one of the main performance limitations of Hybrid-DBT: the lack of speculative execution. We study how it is possible to use memory dependency speculation during the DBT process. Our approach enables fine-grained speculation optimizations thanks to a combination of hardware and software mechanisms. Our results show that our approach leads to a geo-mean speed-up of 10% at the price of a 7% area overhead. In [49], we summarize the current state of the Hybrid-DBT project and display our last results about the performance and the energy efficiency of the system. The experimental results presented here show that, for compute-intensive benchmarks, Hybrid-DBT can deliver the same performance level than a 3-issue OoO core, while consuming three times less energy. Finally, in [34], we investigate security issues caused by the use of speculation in DBT-based systems. We demonstrate that, even if those systems use in-order micro-architectures, the DBT layer optimizes binaries and speculates on the outcome of some branches, leading to security issues similar to the Spectre vulnerability. We demonstrate that both the NVidia Denver architecture and the Hybrid-DBT platform are subject to such vulnerability. However, we also demonstrate that those systems can easily be patched, as the DBT is done in software and has fine-grained control over the optimization process.

6.1.3. What You Simulate Is What You Synthesize: Designing a Processor Core from C++ Specifications

Participants: Simon Rokicki, Davide Pala, Joseph Paturel, Olivier Sentieys.

Designing the hardware of a processor core as well as its verification flow from a single high-level specification would provide great advantages in terms of productivity and maintainability. In [32] (a preliminary version also in [42]), we highlight the gain of starting from a unique high-level synthesis and simulation C++ model to design a processor core implementing the RISC-V Instruction Set Architecture (ISA). The specification code is used to generate both the hardware target design through High-Level Synthesis as well as a fast and cycle-accurate bit-accurate simulator of the latter through software compilation. The object oriented nature of C++ greatly improves the readability and flexibility of the design description compared to classical HDL-based implementations. Therefore, the processor model can easily be modified, expanded and verified using standard software development methodologies. The main challenge is to deal with C++ based synthesizable specifications of core and uncore components, cache memory hierarchy, and synchronization. In particular, the research question is how to specify such parallel computing pipelines with high-level synthesis technology and to demonstrate that there is a potential high gain in design time without jeopardizing performance and cost. Our experiments demonstrate that the core frequency and area of the generated hardware are comparable to existing RTL implementations.

6.1.4. Accelerating Itemset Sampling on FPGA

Participants: Mael Gueguen, Olivier Sentieys.

Finding recurrent patterns within a data stream is important for fields as diverse as cybersecurity or e-commerce. This requires to use pattern mining techniques. However, pattern mining suffers from two issues. The first one, known as ”pattern explosion”, comes from the large combinatorial space explored and is the result of too many patterns outputted to be analyzed. Recent techniques called output space sampling solve this problem by outputting only a sampled set of all the results, with a target size provided by the user. The second issue is that most algorithms are designed to operate on static datasets or low throughput streams. In [24], we propose a contribution to tackle both issues, by designing an FPGA accelerator for pattern mining with output space sampling. We show that our accelerator can outperform a state-of-the-art implementation on a server class CPU using a modest FPGA product. This work is done in collaboration with A. Termier from the Lacodam team at Inria.

6.1.5. Hardware Accelerated Simulation of Heterogeneous Platforms

Participants: Minh Thanh Cong, François Charot, Steven Derrien.

When considering designing heterogeneous multicore platforms, the number of possible design combinations leads to a huge design space, with subtle trade-offs and design interactions. To reason about what design is best for a given target application requires detailed simulation of many different possible solutions. Simulation frameworks exist (such as gem5) and are commonly used to carry out these simulations. Unfortunately, these are purely software-based approaches and they do not allow a real exploration of the design space. Moreover, they do not really support highly heterogeneous multicore architectures. These limitations motivate the use of hardware to accelerate the simulation of heterogeneous multicore, and in particular of FPGA components. We study an approach for designing such systems based on performance models through combining accelerator and processor core models. These models are implemented in the HAsim/LEAP infrastructure. In [22], we propose a methodology for building performance models of accelerators and describe the defined design flow.

6.1.6. Fault-Tolerant Scheduling onto Multicore embedded Systems

Participants: Emmanuel Casseau, Minyu Cui, Petr Dobias, Lei Mo, Angeliki Kritikakou.

Demand on multiprocessor systems for high performance and low energy consumption still increases in order to satisfy our requirements to perform more and more complex computations. Moreover, the transistor size gets smaller and their operating voltage is lower, which goes hand in glove with higher susceptibility to system failure. In order to ensure system functionality, it is necessary to conceive fault-tolerant systems. Temporal and/or spatial redundancy is currently used to tackle this issue. Actually, multiprocessor platforms can be less vulnerable when one processor is faulty because other processors can take over its scheduled tasks. In this context, we investigate how to map and schedule tasks onto homogeneous faulty processors.

We consider two approaches. The first approach deals with task mapping onto processors at compile time. Our goal is to guarantee both reliability and hard real-time constraints with low-energy consumption. Task duplication is assessed and duplication is performed if expected reliability of a task is not met. This work concurrently decides duplication of tasks, the task execution frequency and task allocation to minimize the energy consumption of a multicore platform with Dynamic Voltage and Frequency Scaling (DVFS) capabilities. The problem is initially formulated as Integer Non-Linear Programming and equivalently transformed to a Mixed Integer Linear Programming problem to be optimally solved. The proposed approach provides a good trade-off between energy consumption and reliability. The second approach deals with mapping and scheduling tasks at runtime. The application context is CubeSats. CubeSats operate in harsh space environment and they are exposed to charged particles and radiations, which cause transient faults. To make CubeSats fault tolerant, we propose to take advantage of their multicore architecture. We propose two online algorithms, which schedule all tasks on board of a CubeSat, detect faults and take appropriate measures (based on task replication) in order to deliver correct results. The first algorithm considers all tasks as aperiodic tasks and the second one treats them as aperiodic or periodic tasks. Their performances vary, particularly when the number of processors is low, and a choice is subject to a trade-off between the rejection rate and the energy consumption. This work is done in collaboration with Oliver Sinnen, PARC Lab., the University of Auckland.

6.1.7. Run-Time Management on Multicore Platforms

Participant: Angeliki Kritikakou.

In time-critical systems, run-time adaptation is required to improve the performance of time-triggered execution, derived based on Worst-Case Execution Time (WCET) of tasks. By improving performance, the systems can provide higher Quality-of-Service, in safety-critical systems, or execute other best-effort applications, in mixed-critical systems. To achieve this goal, we propose a parallel interference-sensitive run-time adaptation mechanism that enables a fine-grained synchronisation among cores [37]. Since the run-time adaptation of offline solutions can potentially violate the timing guarantees, we present the Response-Time Analysis (RTA) of the proposed mechanism showing that the system execution is free of timing-anomalies. The RTA takes into account the timing behavior of the proposed mechanism and its associated WCET. To support our contribution, we evaluate the behavior and the scalability of the proposed approach for different application types and execution configurations on the 8-core Texas Instruments TMS320C6678 platform. The obtained results show significant performance improvement compared to state-of-the-art centralized approaches.

6.1.8. Energy Constrained and Real-Time Scheduling and Assignment on Multicores

Participants: Olivier Sentieys, Angeliki Kritikakou, Lei Mo.

Asymmetric Multicore Processors (AMP) are a very promising architecture to deal efficiently with the wide diversity of applications. In real-time application domains, in-time approximated results are preferred to accurate – but too late – results. In [28], we propose a deployment approach that exploits the heterogeneity provided by AMP architectures and the approximation tolerance provided by the applications, so as to increase as much as possible the quality of the results under given energy and timing constraints. Initially, an optimal approach is proposed based on the problem linearization and decomposition. Then, a heuristic approach is developed based on iteration relaxation of the optimal version. The obtained results show 16.3% reduction in the computation time for the optimal approach compared to conventional optimal approaches. The proposed heuristic approach is about 100 times faster at the cost of a 29.8% QoS degradation in comparison with the optimal solution.

6.1.9. Real-Time Energy-Constrained Scheduling in Wireless Sensor and Actuator Networks

Participants: Angeliki Kritikakou, Lei Mo.

Cyber-Physical Systems (CPS), as a particular case of distributed systems, raise new challenges, because of the heterogeneity and other properties traditionally associated with Wireless Sensor and Actuator Networks (WSAN), including shared sensing, acting and real-time computing. In CPS, mobile actuators can enhance system's flexibility and scalability, but at the same time incur complex couplings in the scheduling and controlling of the actuators. In [19], we propose a novel event-driven method aiming at satisfying a required

level of control accuracy and saving energy consumption of the actuators, while guaranteeing a bounded action delay. We formulate a joint-design problem of both actuator scheduling and output control. To solve this problem, we propose a two-step optimization method. In the first step, the problem of actuator scheduling and action time allocation is decomposed into two subproblems. They are solved iteratively by utilizing the solution of one in the other. The convergence of this iterative algorithm is proved. In the second step, an on-line method is proposed to estimate the error and adjust the outputs of the actuators accordingly. Through simulations and experiments,we demonstrate the effectiveness of the proposed method. In addition, many of the real-time tasks of CPS can be executed in an imprecise way. Such systems accept an approximate result as long as the baseline Quality-of-Service (QoS) is satisfied and they can execute more computations to yield better results, if more system resources are available. These systems are typically considered under the Imprecise Computation (IC) model, achieving a better tradeoff between QoS and limited system resources. However, determining a QoS-aware mapping of these real-time IC-tasks onto the nodes of a CPS creates a set of interesting problems. In [18], we firstly propose a mathematical model to capture the dependency, energy and real-time constraints of IC-tasks, as well as the sensing, acting, and routing in the CPS. The problem is formulated as a Mixed-Integer Non-Linear Programming (MINLP) due to the complex nature of the problem. Secondly, to efficiently solve this problem, we provide a linearization method that results in a Mixed-Integer Linear Programming (MILP) formulation of our original problem. Finally, we decompose the transformed problem into a task allocation subproblem and a task adjustment subproblem, and, then, we find the optimal solution based on subproblem iteration. Through the simulations, we demonstrate the effectiveness of the proposed method. Last, but not least, wireless charging can provide dynamic power supply for CPS. Such systems are typically considered under the scenario of Wireless Rechargeable Sensor Networks (WRSNs). With the use of Mobile Chargers (MCs), the flexibility of WRSNs is further enhanced. However, the use of MCs poses several challenges during the system design. The coordination process has to simultaneously optimize the scheduling, the moving time and the charging time of multiple MCs, under limited system resources (e.g., time and energy). Efficient methods that jointly solve these challenges are generally lacking in the literature. In [17], we address the multiple MCs coordination problem under multiple system requirements. Firstly, we aim at minimizing the energy consumption of MCs, guaranteeing that every sensor will not run out of energy. We formulate the multiple MCs coordination problem as a mixed-integer linear programming and derive a set of desired network properties. Secondly, we propose a novel decomposition method to optimally solve the problem, as well as to reduce the computation time. Our approach divides the problem into a subproblem for the MC scheduling and a subproblem for the MC moving time and charging time, and solves them iteratively by utilizing the solution of one into the other. The convergence of the proposed method is analyzed theoretically. Simulation results demonstrate the effectiveness and scalability of the proposed method in terms of solution quality and computation time.

6.1.10. Fault-Tolerant Microarchitectures

Participants: Joseph Paturel, Angeliki Kritikakou, Olivier Sentieys.

As transistors scale down, processors are more vulnerable to radiation that can cause multiple transient faults in function units. Rather than excluding these units from execution, performance overhead of VLIW processors can be reduced when fault-free components of these affected units are still used. In [30], the function units are enhanced with coarse-grained fault detectors. A re-scheduling of the instructions is performed at run-time to use not only the healthy function units, but also the fault-free components of the faulty function units. The scheduling window of the proposed mechanism covers two instruction bundles, which makes it suitable to explore mitigation solutions in the current and in the next instruction execution. Experiments show that the proposed approach can mitigate a large number of faults with low performance and area overheads. In addition, technology scaling can cause transient faults with long duration. In this case, the affected function unit is usually considered as faulty and is not further used. To reduce this performance degradation, we proposed a hardware mechanism to (i) detect the faults that are still active during execution and (ii) re-schedule the instructions to use the fault-free components of the affected function units [31]. When the fault faints, the affected function unit components can be reused. The scheduling window of the proposed mechanism is two instruction bundles being able to exploit function units of both the current and the next instruction execution.

The results show multiple long-duration fault mitigation can be achieved with low performance, area, and power overhead.

Simulation-based fault injection is commonly used to estimate system vulnerability. Existing approaches either partially model the studied system's fault masking capabilities, losing accuracy, or require prohibitive estimation times. Our work proposes a vulnerability analysis approach that combines gate-level fault injection with microarchitecture-level Cycle-Accurate and Bit-Accurate simulation, achieving low estimation time. Faults both in sequential and combinational logic are considered and fault masking is modeled at gate-level, microarchitecture-level and application-level, maintaining accuracy. Our case-study is a RISC-V processor. Obtained results show a more than 8% reduction in masked errors, increasing more than 55% system failures compared to standard fault injection approaches. This work is currently under review.

6.1.11. Fault-Tolerant Networks-on-Chip

Participants: Romain Mercier, Cédric Killian, Angeliki Kritikakou, Daniel Chillet.

Network-on-Chip has become the main interconnect in the multicore/manycore era since the beginning of this decade. However, these systems become more sensitive to faults due to transistor shrinking size. In parallel, approximate computing appears as a new computation model for applications since several years. The main characteristic of these applications is to support the approximation of data, both for computations and for communications. To exploit this specific application property, we develop a fault-tolerant NoC to reduce the impact of faults on the data communications. To address this problem, we consider multiple permanent faults on router which cannot be managed by Error-Correcting Codes (ECCs) and we propose a bit-shuffling method to reduce the impact of faults on Most Significant Bits (MSBs), hence permanent faults only impact Low Significant Bits (LSBs) instead of MSBs reducing the errors impact. We evaluated the proposed method for data mining benchmark and we show that our proposal can lead to 73.04% reduction on the clustering error rate and 84.64% reduction on the mean centroid Mean Square Error (MSE) for 3-bit permanent faults which affect MSBs on 32-bit words with a limited area cost. This work is currently under review for an international conference.

6.1.12. Improving the Reliability of Wireless Network-on-Chip (WiNoC)

Participants: Joel Ortiz Sosa, Olivier Sentieys, Cédric Killian.

Wireless Network-on-Chip (WiNoC) is one of the most promising solutions to overcome multi-hop latency and high power consumption of modern many/multi core System-on-Chip (SoC). However, standard WiNoC approaches are vulnerable to multi-path interference introduced by on-chip physical structures. To overcome such parasitic phenomenon, we first proposed a Time-Diversity Scheme (TDS) to enhance the reliability of on-chip wireless links using a realistic wireless channel model. We then proposed an adaptive digital transceiver, which enhances communication reliability under different wireless channel configurations in [39]. Based on the same realistic channel model, we investigated the impact of using some channel correction techniques. Experimental results show that our approach significantly improves Bit Error Rate (BER) under different wireless channel configurations. Moreover, our transceiver is designed to be adaptive, which allows for wireless communication links to be established in conditions where this would not be possible for standard transceiver architectures. The proposed architecture, designed using a 28-nm FDSOI technology, consumes only 3.27 mW for a data rate of 10 Gbit/s and has a very small area footprint. We also proposed a low-power, high-speed, multi-carrier reconfigurable transceiver based on Frequency Division Multiplexing (FDM) to ensure data transfer in future Wireless NoCs in [38]. The proposed transceiver supports a medium access control method to sustain unicast, broadcast and multicast communication patterns, providing dynamic data exchange among wireless nodes. Designed using a 28-nm FDSOI technology, the transceiver only consumes 2.37 mW and 4.82 mW in unicast/broadcast and multicast modes, respectively, with an area footprint of 0.0138 mm^2.

6.1.13. Error Mitigation in Nanophotonic Interconnect

Participants: Jaechul Lee, Cédric Killian, Daniel Chillet.

The energy consumption of manycore is dominated by data movements, which calls for energy-efficient and high-bandwidth interconnects. Integrated optics is promising technology to overcome the bandwidth limitations of electrical interconnects. However, it suffers from high power overhead related to low efficiency lasers, which calls for the use of approximate communications for error tolerant applications. In this context, in [26] we investigate the design of an Optical NoC supporting the transmission of approximate data. For this purpose, the least significant bits of floating point numbers are transmitted with low power optical signals. A transmission model allows estimating the laser power according to the targeted BER and a micro-architecture allows configuring, at run-time, the number of approximated bits and the laser output powers. Simulation results show that, compared to an interconnect involving only robust communications, approximations in the optical transmissions lead to a laser power reduction up to 42% for image processing application with a limited degradation at the application level.

6.2. Compilation and Synthesis for Reconfigurable Platform

6.2.1. Compile Time Simplification of Sparse Matrix Code Dependences

Participant: Tomofumi Yuki.

In [29], we developed a combined compile-time and runtime loop-carried dependence analysis of sparse matrix codes and evaluated its performance in the context of wavefront parallellism. Sparse computations incorporate indirect memory accesses such as x[col[j]] whose memory locations cannot be determined until runtime. The key contributions are two compile-time techniques for significantly reducing the overhead of runtime dependence testing: (1) identifying new equality constraints that result in more efficient runtime inspectors, and (2) identifying subset relations between dependence constraints such that one dependence test subsumes another one that is therefore eliminated. New equality constraints discovery is enabled by taking advantage of domain-specific knowledge about index arrays, such as col[j]. These simplifications lead to automatically-generated inspectors that make it practical to parallelize such computations. We analyze our simplification methods for a collection of seven sparse computations. The evaluation shows our methods reduce the complexity of the runtime inspectors significantly. Experimental results for a collection of five large matrices show parallel speedups ranging from 2x to more than 8x running on a 8-core CPU.

6.2.2. Study of Polynomial Scheduling

Participant: Tomofumi Yuki.

We have studied the Handelman's theorem used for polynomial scheduling, which resembles the Farkas' lemma for affine scheduling. Theorems from real algebraic geometry and polynomial optimization show that some polynomials have Handelman representations when they are non-negative on a domain, instead of strictly positive as stated in Handelman's theorem. The global minimizers of a polynomial must be at the boundaries of the domain to have such a representation with finite bounds on the degree of monomials. This creates discrepancies in terms of polynomials included in the exploration space with a fixed bound on the monomial degree. Our findings give an explanation to our failed attempt to apply polynomial scheduling to Index-Set Splitting: we were precisely trying to find polynomials with global minimizers at the interior of a domain.

6.2.3. Optimizing and Parallelizing compilers for Time-Critical Systems

Participant: Steven Derrien.

6.2.3.1. Contentions-Aware Task-Level Parallelization

Accurate WCET analysis for multicores is challenging due to concurrent accesses to shared resources, such as communication through bus or Network on Chip (NoC). Current WCET techniques either produce pessimistic WCET estimates or preclude conflicts by constraining the execution, at the price of a significant hardware under-utilization. Most existing techniques are also restricted to independent tasks, whereas real-time workloads will probably evolve toward parallel programs. The WCET behavior of such parallel programs is even more challenging to analyze because they consist of *dependent* tasks interacting through complex synchronization/communication mechanisms. In [36], we propose a scheduling technique that jointly selects

Scratchpad Memory (SPM) contents off-line, in such a way that the cost of SPM loading/unloading is hidden. Communications are fragmented to augment hiding possibilities. Experimental results show the effectiveness of the proposed technique on streaming applications and synthetic task-graphs. The overlapping of communications with computations allows the length of generated schedules to be reduced by 4% on average on streaming applications, with a maximum of 16%, and by 8% on average for synthetic task graphs. We further show on a case study that generated schedules can be implemented with low overhead on a predictable multicore architecture (Kalray MPPA).

6.2.3.2. WCET-Aware Parallelization of Model-Based Applications for Multicores

Parallel architectures are nowadays increasingly used in embedded time-critical systems. The Argo H2020 project provides a programming paradigm and associated tool flow to exploit the full potential of architectures in terms of development productivity, time-to-market, exploitation of the platform computing power and guaranteed real-time performance. The Argo toolchain operates on Scilab and XCoS inputs, and targets ScratchPad Memory (SPM)-based multicores. Data-layout and loop transformations play a key role in this flow as they improve SPM efficiency and reduce the number of accesses to shared main memory. In [20] we present the overall results of the project, a compiler tool-flow for automated parallelization of model-based real-time software, which addresses the shortcomings of multi-core architectures in real-time systems. The flow is demonstrated using a model-based Terrain Awareness and Warning Systems (TAWS) and an edge detection algorithm from the image-processing domain. Model-based applications are first transformed into real-time C code and from there into a well-predictable parallel C program. Tight bounds for the Worst-Case Execution Time (WCET) of the parallelized program can be determined using an integrated multicore WCET analysis. Thanks to the use of an architecture description language, the general approach is applicable to a wider range of target platforms. An experimental evaluation for a research architecture with network-on-chip (NoC) interconnect shows that the parallel WCET of the TAWS application can be improved by factor 1.77 using the presented compiler tools.

6.2.3.3. WCET oriented Iterative compilation

Static Worst-Case Execution Time (WCET) estimation techniques operate upon the binary code of a program in order to provide the necessary input for schedulability analysis techniques. Compilers used to generate this binary code include tens of optimizations, that can radically change the flow information of the program. Such information is hard to maintain across optimization passes and may render automatic extraction of important flow information, such as loop bounds, impossible. Thus, compiler optimizations, especially the sophisticated optimizations of mainstream compilers, are typically avoided. In this work, published in [23], we explore for the first time iterative-compilation techniques that reconcile compiler optimizations and static WCET estimation. We propose a novel learning technique that selects sequences of optimizations that minimize the WCET estimate of a given program. We experimentally evaluate the proposed technique using an industrial WCET estimation tool (AbsInt aiT) over a set of 46 benchmarks from four different benchmarks suites, including reference WCET benchmark applications, image processing kernels and telecommunication applications. Experimental results show that WCET estimates are reduced on average by 20.3% using the proposed technique,as compared to the best compiler optimization level applicable.

6.2.4. Towards Generic and Scalable Word-Length Optimization

Participants: Van-Phu Ha, Tomofumi Yuki, Olivier Sentieys.

Fixed-Point arithmetic is widely used for implementing Digital Signal Processing (DSP) systems on electronic devices. Since initial specifications are often written using floating-point arithmetic, conversion to fixed-point is a recurring step in hardware design. The primary objective of this conversion is to minimize the cost (energy and/or area) while maintaining an acceptable level of quality at the output. In Word-Length Optimization (WLO), each variable/operator may be assigned a different fixed-point encoding, which means that the design space grows exponentially as the number of variables increases. This is especially true when targeting hardware accelerators implemented in FPGA or ASIC. Thus, most approaches for WLO involve heuristic search algorithms. In [25] (a preliminary version also in [41]), we propose a method to improve the scalability of Word-Length Optimization (WLO) for large applications that use complex quality metrics such as Structural

Similarity (SSIM). The input application is decomposed into smaller kernels to avoid uncontrolled explosion of the exploration time, which is known as noise budgeting. The main challenge addressed in this paper is how to allocate noise budgets to each kernel. This requires capturing the interactions across kernels. The main idea is to characterize the impact of approximating each kernel on accuracy/cost through simulation and regression. Our approach improves the scalability while finding better solutions for Image Signal Processor pipeline.

In [27], we propose an analytical approach to study the impact of floating-point (FlP) precision variation on the square root operation, in terms of computational accuracy and performance gain. We estimate the round-off error resulting from reduced precision. We also inspect the Newton Raphson algorithm used to approximate the square root in order to bound the error caused by algorithmic deviation. Consequently, the implementation of the square root can be optimized by fittingly adjusting its number of iterations with respect to any given FlP precision specification, without the need for long simulation times. We evaluate our error analysis of the square root operation as part of approximating a classic data clustering algorithm known as K-means, for the purpose of reducing its energy footprint. We compare the resulting inexact K-means to its exact counterpart, in the context of color quantization, in terms of energy gain and quality of the output. The experimental results show that energy savings could be achieved without penalizing the quality of the output (e.g., up to 41.87% of energy gain for an output quality, measured using structural similarity, within a range of [0.95,1]).

6.2.5. Optimized Implementations of Constant Multipliers for FPGAs

Participant: Silviu-Ioan Filip.

The multiplication by a constant is a frequently used arithmetic operation. To implement it on Field Programmable Gate Arrays(FPGAs), the state of the art offers two completely different methods: one relying on bit shifts and additions/subtractions, and another one using look-up tables and additions. So far, it was unclear which method performs best for a given constant and input/output data types. The main contribution of the work published in [40] is a thorough comparison of both methods in the main application contexts of constant multiplication: filters, signal-processing transforms, and elementary functions. Most of the previous state of the art addresses multiplication by an integer constant. This work shows that, in most of these application contexts, a formulation of the problem as the multiplication by a real constant allows for more efficient architectures. Another contribution is a novel extension of the shift-and-add method to real constants. For that, an integer linear programming (ILP) formulation is proposed, which truncates each component in the shift-and-add network to a minimum necessary word size that is aligned with the approximation error of the coefficient. All methods are implemented within the open-source FloPoCo framework.

6.2.6. Optimal Multiplierless FIR Filter Design

Participant: Silviu-Ioan Filip.

The hardware optimization of direct form finite impulse response (FIR) filters has been a topic of research for the better part of the last four decades and is still garnering significant research and industry interest. In [48], we present two novel optimization methods based on integer linear programming (ILP) that minimize the number of adders used to implement a direct/transposed FIR filter adhering to a given frequency specification. The proposed algorithms work by either fixing the number of adders used to implement the products (multiplier block adders) or by bounding the adder depth (AD) used for these products. The latter can be used to design filters with minimal AD for low power applications. In contrast to previous multiplierless FIR approaches, the methods introduced here ensure adder count optimality. To demonstrate their effectiveness, we perform several experiments using established design problems from the literature, showing superior results.

6.2.7. Application-specific arithmetic in high-level synthesis tools

Participant: Steven Derrien.

In [50], we have shown that the use of non-conventional implementation for floating-point arithmetic can bring significant benefits when used in the context of High-Level Synthesis. We are currently building on these preliminary results to show that it is possible to implement accelerators using exact floating-point arithmetic for similar performance/area cost than standard floating-point operators implementations. Our approach builds on Kulish's approach to implement floating-point adders, and targets dense Matrix Products kernels (GEM3 like) accelerators on FPGAs.

6.3. Applications

6.3.1. SmartSense

Participants: Nicolas Roux, Olivier Sentieys.

Developing smarter and greener buildings has been an expanding field of research over the last decades. One of the essential requirements for energy utilities is the knowledge of power consumption patterns at the single-appliance level. To estimate these patterns without using an individual power meter for each appliance, Non-Intrusive Load Monitoring (NILM) consists in disaggregating electrical loads by examining the appliance specific power consumption signature within the aggregated load single measurement. Therefore, the method is considered non-intrusive since the data are collected from a single electrical panel outside of the monitored building. Thus, NILM has been a very active field of research with renewed interest over the last years.

Therefore, knowing the plug-level power consumption of each appliance in a building can lead to drastic savings in energy consumption. In [35], we have addressed the issue of NILM inaccuracy in the context of industrial or commercial buildings, by combining data from a low-cost, general-purpose, wireless sensor network. We have proposed a novel approach based on a simplex solver to estimate the power load values of the steady states on sliding windows of data with varying size. We have shown the principle of the approach and demonstrated its interest, limited complexity, and ease of use.

CELTIQUE Project-Team

4. New Results

4.1. Compiling Sandboxes:Formally Verified Software Fault Isolation

Participants: Frédéric Besson, Sandrine Blazy, Alexandre Dang, Thomas Jensen.

Software Fault Isolation (SFI) is a security-enhancing pro- gram transformation for instrumenting an untrusted binary module so that it runs inside a dedicated isolated address space, called a sandbox. To ensure that the untrusted module cannot escape its sandbox, existing approaches such as Google's Native Client rely on a binary verifier to check that all memory accesses are within the sandbox. Instead of rely- ing on a posteriori verification, we design, implement and prove correct a program instrumentation phase as part of the formally verified compiler CompCert that enforces a sandboxing security property a priori. This eliminates the need for a binary verifier and, instead, leverages the soundness proof of the compiler to prove the security of the sandbox- ing transformation. The technical contributions are a novel sandboxing transformation that has a well-defined C semantics and which supports arbitrary function pointers, and a formally verified C compiler that im- plements SFI. Experiments show that our formally verified technique is a competitive way of implementing SFI [6].

4.2. Information-Flow Preservation in Compiler Optimisations

Participants: Frédéric Besson, Alexandre Dang, Thomas Jensen.

Correct compilers perform program transformations preserving input/output behaviours of programs. Though mandatory, correctness is not sufficient to prevent program optimisations from introducing information-flow leaks that would make the target program more vulnerable to side-channel at- tacks than the source program. To tackle this problem, we propose a notion of Information-Flow Preserving (IFP) program transformation which ensures that a target program is no more vulnerable to passive side-channel attacks than a source program. To protect against a wide range of attacks, we model an attacker who is granted arbitrary memory accesses for a pre-defined set of observation points. We have proposed a compositional proof principle for proving that a transformation is IFP. Using this principle, we show how a translation validation technique can be used to automatically verify and even close information-flow leaks introduced by standard compiler passes such as dead- store elimination and register allocation. The technique has been experimentally validated on the CompCert C compiler [7].

4.3. Formalization of Higher-Order Process Calculi

Participants: Guillaume Ambal, Alan Schmitt.

Guillaume Amabal and Alan Schmitt, in collaboration with Sergueï Lenglet, have continued exploring how to formalize HOπ in Coq, in particular how to deal with the different kinds of binders used in the calculus. We have extended our previous study that compared locally nameless, De Bruijn indices, and nominal binders with an approach based on higher-order abstract syntax. We have discovered that this approach is not as elegant as in other calculi. A journal version is submitted for publication. The Coq scripts can be found at http://passivation.gforge.inria.fr/hopi/.

4.4. Certified Semantics and Analyses for JavaScript

Participants: Samuel Risbourg, Alan Schmitt.

Alan Schmitt and Samuel Risbourg have continued to develop JSExplain, an interpreter for JavaScript that is as close as possible to the specification. The tool is publicly available at https://github.com/jscert/jsexplain. It was presented to the TC39 committee standardizing JavaScript in December to solicit feedback.

4.5. Skeletal Semantics

Participants: Guillaume Ambal, Nathanael Courant, Thomas Jensen, Adam Khayam, Louis Noizet, Vincent Rebiscoul, Alan Schmitt.

The work on skeletal semantics [5], a modular and formal way to describe semantics or programming languages, has intensified during 2019. We have continued to develop `necro`, a tool to manipulate skeletal semantics and generate interpreters in OCaml, mechanized semantics in Coq, and static analyzers. The code is available online (). Several interns and PhD students are also working on skeletal semantics.

Nathanaël Courant has designed a control-flow analyzer for languages written as skeletal semantics. This work is now extended by Vincent Rebiscoul to certify the analyzer.

Louis Noizet is studying the formalization in Coq of natural semantics from skeletal semantics. To this end, he extended the necro tool to automatically generate a Coq formalization. Louis is also very involved in the maintenance of necro.

Guillaume Ambal is studying the language features that can be captured using skeletal semantics, focusing on concurrency and distribution. In this setting, he is building an approach to automatically derive a small-step semantics form a big-step one.

Adam Khayam is writing a formal semantics of the Hop multitier language, an extension of JavaScript to write web applications. As a first step, he is writing a skeletal semantics of JavaScript to validate that our approach scale for complex and sizable semantics.

4.6. Static analyses for proofs of programs

Participants: Oana Andreescu, Thomas Jensen, Stéphane Lescuyer, Benoît Montagu.

Thomas Jensen together with three industrial research engineers Oana Andreescu, Stéphane Lescuyer, and Benoît Montagu, worked on the development of static analyses that help reduce the manual proof effort that is needed to formally verify programs.

They improved the correlation analysis that Oana Andreescu introduced in her Phd thesis, by designing a novel abstract domain. They verified in Coq its semantic properties, and evaluated their approach on an industrial micro-kernel developed at *Prove&Run*. They showed that the technique could reduce the proof burden by two thirds [1].

4.7. Constant-time verification by compilation and static analysis

Participants: Sandrine Blazy, David Pichardie, Alix Trieu.

To protect their implementations, cryptographers follow a very strict programming discipline called constant-time programming. They avoid branchings controlled by secret data as an attacker could use timing attacks, which are a broad class of side-channel attacks that measure different execution times of a program in order to infer some of its secret values. Several real-world secure C libraries such as NaCl, mbedTLS, or Open Quantum Safe, follow this discipline. We propose an advanced static analysis, based on state-of-the-art techniques from abstract interpretation, to report time leakage during programming. To that purpose, we analyze source C programs and use full context-sensitive and arithmetic-aware alias analyses to track the tainted flows. We give semantic evidences of the correctness of our approach on a core language. We also present a prototype implementation for C programs that is based on the CompCert compiler toolchain and its companion Verasco static analyzer. We present verification results on various real-world constant-time programs and report on a successful verification of a challenging SHA-256 implementation that was out of scope of previous tool-assisted approaches. This work has been published in [4] as an extended version of [12].

CIDRE Project-Team

6. New Results

6.1. Axis 1 : Attack comprehension

6.1.1. Fault injection

Electromagnetic injection is a non-invasive way to attack a chip. The large number of parameters that require to be properly tuned for such an attack limits its efficiency. In [30] we propose several ways to improve the success rate of fault injection by electromagnetic radiation. We show that software execution is altered at targeted instructions if the radiating probe is located above the phase-locked loop device driving the clock tree. We identify the phase-locked loop as a sensitive part of the chip. We reduce the preferential location for the electromagnetic injection to a small area in the vicinity of the analog power supply feeding the phase-locked loop. We also explore the influence of the frequency of the injected electromagnetic wave. We compute the optimal fault rate in a bandwidth of $15MHz$, in the upper limit of the chip bandwidth. Our experiments show that for an optimal frequency a precision of $5ns$, we succeed to reach the best fault rate. With this electromagnetic injection technique, the achieved success rate reaches 15 to 20%. Such a fault can be used to retrieve the key of a cryptographic algorithm (for an Advanced Encryption Standard application for example).

6.1.2. Malware analysis

About Android malware analysis, we have started investigations with specific malware that hide their behavior using obfuscation techniques [10]. As these malware are difficult to find in the wild, we have also started to analyze both datasets of the literature and large collection of applications captured from different repositories such as the Play Store. This huge amount of applications to analyze (currently more than 100,000) makes difficult to build reliable experiments [20]. We have designed a new tool, called PyMaO, that helps to orchestrate experiments. This tool is published as an open source tool under GPL v3. We have also revisited the historical datasets of malware of the literature and introduce a more up-to-date malware and goodware dataset [26].

6.1.3. Focus on doxware

A doxware is a particular type of ransomware that threatens to release personal or sensitive data to the public if the user does not pay the ransom. The term comes from the hacker term "doxing," or releasing confidential information over the internet. The only difference between a classical ransomware and a doxware resides in a *valuable files hunting* followed by an exfiltration of these data. In [34], we have explored how an attacker may be able to quickly localized valuable assets of a machine using an analysis of the content and the vocabulary of its files.

6.1.4. Attack scenario reconstruction

In order to supervise the security of a large infrastructure, the administrator deploys multiple sensors and intrusion detection systems on several critical places in the system. It is easier to explain and detect attacks if more events are logged. Starting from a suspicious event (appearing as a log entry), the administrator can start his investigation by manually building the set of previous events that are linked to this event of interest. Accordingly, the administrator attempts to identify links among the logged events in order to retrieve those that correspond to the traces of the attacker's actions in the supervised system; previous work is aimed at building these connections. In practice, however, this type of link is not trivial to define and discover. Hence, there is a real necessity to describe and define formally the semantics of these links in literature. In In order to supervise the security of a large infrastructure, the administrator deploys multiple sensors and intrusion detection systems on several critical places in the system. It is easier to explain and detect attacks if more events are logged. Starting from a suspicious event (appearing as a log entry), the administrator can start

his investigation by manually building the set of previous events that are linked to this event of interest. Accordingly, the administrator attempts to identify links among the logged events in order to retrieve those that correspond to the traces of the attacker's actions in the supervised system; previous work is aimed at building these connections. In practice, however, this type of link is not trivial to define and discover. Hence, there is a real necessity to describe and define formally the semantics of these links in literature. In this paper, a clear definition of this relationship, called contextual event causal dependency, is introduced and proposed. The work presented in this paper aims at defining a formal model that would ideally unify previous work on causal dependencies among heterogeneous events. We define a relationship among events that enables the discovery of all events, which can be considered as the cause (in the past) or the effect (in the future) of an event of interest (e.g., an indicator of compromise, produced by an attacker action). In [36], we have proposed a clear definition of this relationship, called contextual event causal dependency. The work presented in [36] aims at defining a formal model that would ideally unify previous work on causal dependencies among heterogeneous events. We define a relationship among events that enables the discovery of all events, which can be considered as the cause (in the past) or the effect (in the future) of an event of interest (e.g., an indicator of compromise, produced by an attacker action).

6.2. Axis 2 : Attack detection

6.2.1. Vulnerabilities detection in Java

In a prior work, we have focused on adapting a machine-learning tool (ChuckyJava) aiming at automatically detect vulnerabilities in Java. ChuckyJava is able to detect vulnerabilities by performing in two steps: the neighborhood discovery and the anomaly detection. The neighborhood discovery is the ability for the tool to detect method of similar semantics: neighbors. In [25], we mitigate many ChuckyJava's limitations by developing JavaNeighbors that improves the neighborhood discovery. JavaNeighbors represents methods by terms and using a method based on a Natural Language Processing technique, JavaNeighbors computes the distance between all representations of methods. Finally, according to the distance, each method has a neighbor list from the closest to the most distant ones. JavaNeighbors has enabled ChuckyJava to detect vulnerabilities with more accuracy.

6.2.2. Ransomware detection

A ransomware attacks mostly begins with social engineering methods to install payloads on victims' computers, followed by a communication with command and control servers for data exchange. To enable an early detection and thus scale down these attacks, we propose in [35] a detection model based on the collected system and network logs from a computer. The analysis is performed on various ransomware families with a high detection rate. Packet level detection is performed to grant the best use case scenario. This work intends to provide an independent third-party procedure that is able to distinguish between a benign software and a malicious ransomware based on network activity. Furthermore, it is not limited to only identify ransomware but could be utilized to inspect different malware.

6.2.3. Intrusion detection using logs of distributed application

Although security issues are now addressed during the development process of distributed applications, an attack may still affect the provided services or allow access to confidential data. To detect intrusions [22], we consider an anomaly detection mechanism which relies on a model of the monitored application's normal behavior. During a model construction phase, the application is run multiple times to observe some of its correct behaviors. Each gathered trace enables the identification of significant events and their causality relationships, without requiring the existence of a global clock. The constructed model is dual: an automaton plus a list of likely invariants. The redundancy between the two sub-models decreases when generalization techniques are applied on the automaton. Solutions already proposed suffer from scalability issues. In particular, the time needed to build the model is important and its size impacts the duration of the detection phase. The proposed solutions address these problems, while keeping a good accuracy during the detection phase, in terms of false positive and false negative rates. To evaluate them, a real distributed application and several attacks against the service have been considered. One of our goal is to identify redundancies and complementarities between the proposed models.

6.3. Axis 3 : Attack resistance

6.3.1. Attacker Life cycle

We have been witnessing for years the awareness of the existence of a so-called Advanced Persistent Threat (APT). These attacks, regularly target or involving nation-states and large companies, were first defined in 2011. Ad Advanced Persistent Threat: (i) pursues its objectives repeatedly over an extended period of time; (ii) adapts to defenders' efforts to resist it; and (iii) is determined to maintain the level of interaction needed to execute its objectives. In [13], we have proposed a model providing an operational reading of the attackers' lifecycle in a compromised network. This model allows to express possible regressions in the attack and introduces the concept of a waiting state, which is essential for long-term actions. In this article we have also proposed a confrontation between our model and two recent examples of attacks whose progression has been publicly described: the Equifax breach (2017) and the TV5Monde sabotage (2015).

6.3.2. OS-level intrusion survivability

Despite the deployment of preventive security mechanisms to protect the assets and computing platforms of users, intrusions eventually occur. In [17], we have proposed a novel intrusion survivability approach to withstand ongoing intrusions. Our approach relies on an orchestration of fine-grained recovery and per-service responses (e.g., privileges removal). Such an approach may put the system into a degraded mode. This degraded mode prevents attackers to reinfect the system or to achieve their goals if they managed to reinfect it. It maintains the availability of core functions while waiting for patches to be deployed. We devised a cost-sensitive response selection process to ensure that while the service is in a degraded mode, its core functions are still operating. We built a Linux-based prototype and evaluated the effectiveness of our approach against different types of intrusions. The results show that our solution removes the effects of the intrusions, that it can select appropriate responses, and that it allows services to survive when reinfected. In terms of performance overhead, in most cases, we observed a small overhead, except in the rare case of services that write many small files asynchronously in a burst, where we observed a higher but acceptable overhead.

6.3.3. Secure routing in drones swarms

Unmanned aerial vehicle (UAV) applications and development have increased over the past few years as this technology has become more accessible and less expensive. On a single UAV scenario, communication is a keystone to transmit commands and retrieve data from UAV sensors. It is even more critical in swarm where cooperation and inter messaging is fundamental. The communication between the nodes of a swarm is based on a suitable routing algorithm. The routing must allow each node to send messages to each other, by successive hops between different neighbors. A UAV swarm is a particular mobile ad-hoc networks where nodes run independently but form a cooperative communication network. UAV swarm shares com- mon characteristics with VANET (vehicular ad hoc network), sensors network or mobile phone network but also strongly differs on specific points (mobility model, instability, limited infrastructure access). Any computation on a UAV is a permanent trade off between volume, weight and power consumption, with no infrastructure access. In [31], we have proposed a secured routing protocol designed for UAV swarm networks. SEER4US is the first protocol providing integrity of routing messages and authentication of their sender with low energy consumption for battery preservation.

6.3.4. Securing the control flow of smartcard C programs

Results obtained several years ago about securing the control flow of C programs have been extended and published in the journal Computers and Security [7]. This extended version of our work focuses on the formal verification of the introduced countermeasures. We prove that any possible attack that would skip more than one C instruction is detected by our countermeasures. We also extended the experimental results on a benchmark software dedicated to smartcards. This work has been achieved in cooperation with Karine Heydemann from the LIP6 laboratory (Sorbonne Université).

6.3.5. A secure implementation of the replicated state machine

State machine replication (RSM) is today the foundation of many cloud-based highly-available products: it allows some service to be deployed such to guarantee its correct functioning despite possible faults. In RSM, clients issue operation requests to a set of distributed processes implementing the replicated service, that, in turn, run a protocol to decide the order of execution of incoming operations and provide clients with outputs. Faults can be accidental (e.g. a computer crashing due to a loss of power) or have a malicious intent (e.g. a compromised server). Whichever is the chosen fault model, RSM has proven to be a reliable and effective solution for the deployment of dependable services. RSM is usually built on top of a distributed Consensus primitive that is used by processes to agree on the order of execution of requests concurrently issued by clients. The main problem with this approach is that Consensus is impossible to achieve deterministically in a distributed settings if the system is asynchronous and even just a single process may fail by crashing. This led the research community to study and develop alternative solutions based on the relaxation of some of the constraints, to allow agreement to be reached in partially synchronous systems with faulty processes by trading off consistency with availability. An alternative approach consists in imposing constraints on the set of operations that can be issued by clients, i.e. imposing updates that commute. In particular, commutative replicated data types (CRDTs) can be implemented with an RSM approach in asynchronous settings using the monotonic growth of a join semilattice, i.e., a partially ordered set that defines a join (least upper bound) for all element pairs. In [18] we have proposed an algorithm that solves Generalized Lattice Agreement in a Byzantine fault model. To the best of our knowledge this is the first solution for Byzantine lattice agreement that works on any possible lattice, and it is the first work proposing a Byzantine tolerant RSM built on it. The algorithm is wait-free, i.e., every process completes its execution of the algorithm within a bounded number of steps, regardless of the execution of other processes. We have also sketch the main lines of a signature-based version of our algorithms which take advantage of digital signatures to reduce the message complexity to $\mathcal{O}(n)$ per process, when the number f of Byzantine processes verifies $f = \mathcal{O}(1)$.

6.3.6. Blockchain in adversarial environments

We are pursuing our efforts dedicated to the theoretical aspects of blockchains. In particular, we have recently proposed to specify blockchains as a composition of abstract data types all together with a hierarchy of consistency criteria that formally characterizes the histories admissible for distributed programs that use them. Our work is based on an original oracle-based construction that, along with new consistency definitions, captures the eventual convergence process in blockchain systems. This study allows us to focus on the implementability of the presented abstractions and a mapping of representative existing blockchains from both academia and industry in our framework. It is already known that some blockchain implementations solve eventual consistency of an append-only queue using Consensus. However the question about the consistency criterion of blockchains as Bitcoin and Ethereum that technically do not solve Consensus, and their relation with Consensus in general was not studied. We have also proposed a specification of distributed ledger register that matches the Lamport hierarchy from safe to atomic. Moreover, we propose implementations of distributed ledger registers with safe, regular and atomic guaranties in a model of communication specific to distributed ledgers technology that we also formalize. Then, we propose an implementation of a distributed ledger register that satisfies the atomic specification and the k-consistency property that characterizes the permissionless distributed blockchains such as Bitcoin and Ethereum. Preliminary results appear in [41].

In parallel to this work, we have proposed the design of a scalable permissionless blockchain in the proof-of-stake setting. In particular, we use a distributed hash table as a building block to set up randomized shards, and then leverage the sharded architecture to validate blocks in an efficient manner. We combine verifiable Byzantine agreements run by shards of stakeholders and a block validation protocol to guarantee that forks occur with negligible probability. We impose induced churn to make shards robust to eclipse attacks, and we rely on the UTXO coin model to guarantee that any stake-holder action is securely verifiable by anyone. Our protocol works against adaptive adversary, and makes no synchrony assumption beyond what is required for the byzantine agreement. This work has been published in [19].

GALLINETTE Project-Team

6. New Results

6.1. Logical Foundations of Programming Languages

Participants: Esaie Bauer, Rémi Douence, Marie Kerjean, Ambroise Lafont, Maxime Lucas, Étienne Miquey, Guillaume Munch-Maccagnoni, Nicolas Tabareau.

6.1.1. Classical Logic

6.1.1.1. Continuation-and-environment-passing style translations: a focus on call-by-need

The call-by-need evaluation strategy for the λ-calculus is an evaluation strategy that lazily evaluates arguments only if needed, and if so, shares computations across all places where it is needed. To implement this evaluation strategy, abstract machines require some form of global environment. While abstract machines usually lead to a better understanding of the flow of control during the execution, easing in particular the definition of continuation-passing style translations, the case of machines with global environments turns out to be much more subtle. The main purpose of [21] is to understand how to type a continuation-and-environment-passing style translations, that it to say how to soundly translate a classical calculus with environment into a calculus that does not have these features. To this end, we focus on a sequent calculus presentation of a call-by-need λ-calculus with classical control for which Ariola et. al already defined an untyped translation and which we equipped with a system of simple types in a previous paper. We present here a type system for the target language of their translation, which highlights a variant of Kripke forcing related to the environment-passing part of the translation. Finally, we show that our construction naturally handles the cases of call-by-name and call-by-value calculi with environment, encompassing in particular the Milner Abstract Machine, a machine with global environments for the call-by-name λ-calculus.

6.1.1.2. Revisiting the duality of computation: an algebraic analysis of classical realizability models

In an impressive series of papers, Krivine showed at the edge of the last decade how classical realizability provides a surprising technique to build models for classical theories. In particular, he proved that classical realizability subsumes Cohen's forcing, and even more, gives rise to unexpected models of set theories. Pursuing the algebraic analysis of these models that was first undertaken by Streicher, Miquel recently proposed to lay the algebraic foundation of classical realizability and forcing within new structures which he called implicative algebras. These structures are a generalization of Boolean algebras based on an internal law representing the implication. Notably, implicative algebras allow for the adequate interpretation of both programs (i.e. proofs) and their types (i.e. formulas) in the same structure. The very definition of implicative algebras takes position on a presentation of logic through universal quantification and the implication and, computationally, relies on the call-by-name λ-calculus. In [13], we investigate the relevance of this choice, by introducing two similar structures. On the one hand, we define disjunctive algebras, which rely on internal laws for the negation and the disjunction and which we show to be particular cases of implicative algebras. On the other hand, we introduce conjunctive algebras, which rather put the focus on conjunctions and on the call-by-value evaluation strategy. We finally show how disjunctive and conjunctive algebras algebraically reflect the well-known duality of computation between call-by-name and call-by-value.

6.1.2. Models of programming languages mixing effects and resources

6.1.2.1. Efficient deconstruction with typed pointer reversal

Building on the connection between resource management in systems programming and ordered logic we established previously, we investigate a pervasive issue in the languages C++ and Rust whereby compiler-generated clean-up functions cause a stack overflow on deep structures. In [17], we show how to generate clean-up algorithms that run in constant time and space for a broad class of ordered algebraic datatypes such as ones that can be found in C++ and Rust or in future extensions of functional programming languages with first-class resources.

6.1.2.2. Resource safety in OCaml

Building on our investigations for a resource-management model for OCaml, we have proposed several preliminary improvements to the OCaml language. We contributed to the design and implementation of new resource management primitives (PRs #2118, #8962), resource-safe C APIs (PRs #8993, #8997, #9037), and core runtime capabilities (PR #8961). (#2118 has been merged into OCaml 4.08 and #8993 and #9037 have been merged into OCaml 4.10.)

We continued to interact with L. White and S. Dolan (Jane Street), on the design of resource management and exception safety in multicore OCaml.

6.1.3. Syntax and Rewriting Systems

6.1.3.1. Reduction Monads and Their Signatures

In [1], we study reduction monads, which are essentially the same as monads relative to the free functor from sets into multigraphs. Reduction monads account for two aspects of the lambda calculus: on the one hand, in the monadic viewpoint, the lambda calculus is an object equipped with a well-behaved substitution; on the other hand, in the graphical viewpoint, it is an oriented multigraph whose vertices are terms and whose edges witness the reductions between two terms. We study presentations of reduction monads. To this end, we propose a notion of reduction signature. As usual, such a signature plays the role of a virtual presentation, and specifies arities for generating operations-possibly subject to equations-together with arities for generating reduction rules. For each such signature, we define a category of models; any model is, in particular, a reduction monad. If the initial object of this category of models exists, we call it the reduction monad presented (or specified) by the given reduction signature. Our main result identifies a class of reduction signatures which specify a reduction monad in the above sense. We show in the examples that our approach covers several standard variants of the lambda calculus.

6.1.3.2. Modules over monads and operational semantics

[22] is a contribution to the search for efficient and high-level mathematical tools to specify and reason about (abstract) programming languages or calculi. Generalising the reduction monads of Ahrens et al., we introduce operational monads, thus covering new applications such as the-calculus, Positive GSOS specifications, and the big-step, simply-typed, call-by-value-calculus. Finally, we design a notion of signature for operational monads that covers all our examples.

6.1.3.3. Modular specification of monads through higher-order presentations

In their work on second-order equational logic, Fiore and Hur have studied presentations of simply typed languages by generating binding constructions and equations among them. To each pair consisting of a binding signature and a set of equations, they associate a category of 'models', and they give a monadicity result which implies that this category has an initial object, which is the language presented by the pair. In [10], we propose, for the untyped setting, a variant of their approach where monads and modules over them are the central notions. More precisely, we study, for monads over sets, presentations by generating ('higher-order') operations and equations among them. We consider a notion of 2-signature which allows to specify a monad with a family of binding operations subject to a family of equations, as is the case for the paradigmatic example of the lambda calculus, specified by its two standard constructions (application and abstraction) subject to β- and η-equalities. Such a 2-signature is hence a pair (Σ,E) of a binding signature Σ and a family E of equations for Σ. This notion of 2-signature has been introduced earlier by Ahrens in a slightly different context. We associate, to each 2-signature (Σ,E), a category of 'models of (Σ,E); and we say that a 2-signature is 'effective' if this category has an initial object; the monad underlying this (essentially unique) object is the 'monad specified by the 2-signature'. Not every 2-signature is effective; we identify a class of 2-signatures, which we call 'algebraic', that are effective. Importantly, our 2-signatures together with their models enjoy 'modularity': when we glue (algebraic) 2-signatures together, their initial models are glued accordingly. We provide a computer formalization for our main results.

6.1.3.4. The Diamond Lemma for non-terminating rewriting systems

In [16], we study the confluence property for rewriting systems whose underlying set of terms admits a vector space structure. For that, we use deterministic reduction strategies. These strategies are based on the choice of standard reductions applied to basis elements. We provide a sufficient condition of confluence in terms of the kernel of the operator which computes standard normal forms. We present a local criterion which enables us to check the confluence property in this framework. We show how this criterion is related to the Diamond Lemma for terminating rewriting systems

6.1.4. Differential Linear Logic

6.1.4.1. Higher-order distributions for differential linear logic

Linear Logic was introduced as the computational counterpart of the algebraic notion of linearity. Differential Linear Logic refines Linear Logic with a proof-theoretical interpretation of the geometrical process of differentiation. In [24], we construct a polarized model of Differential Linear Logic satisfying computational constraints such as an interpretation for higher-order functions, as well as constraints inherited from physics such as a continuous interpretation for spaces. This extends what was done previously by Kerjean for first order Differential Linear Logic without promotion. Concretely, we follow the previous idea of interpreting the exponential of Differential Linear Logic as a space of higher-order distributions with compact-support, and is constructed as an inductive limit of spaces of distributions on Euclidean spaces. We prove that this exponential is endowed with a co-monadic like structure, with the notable exception that it is functorial only on isomorphisms. Interestingly, as previously argued by Ehrhard, this still allows one to interpret differential linear logic without promotion.

6.1.4.2. Chiralities in topological vector spaces

Chiralities are categories introduced by Mellies to account for a game semantics point of view on negation. In [23], [20], we uncover instances of this structure in the theory of topological vector spaces, thus constructing several new polarized models of Multiplicative Linear Logic. These models improve previously known smooth models of Differential Linear Logic, showing the relevance of chiralities to express topological properties of vector spaces. They are the first denotational polarized models of Multiplicative Linear Logic, based on the pre-existing theory of topological vector spaces, in which two distinct sets of formulas, two distinct negations, and two shifts appear naturally.

6.1.5. Distributed Programming

6.1.5.1. Chemical foundations of distributed aspects.

Distributed applications are challenging to program because they have to deal with a plethora of concerns, including synchronisation, locality, replication, security and fault tolerance. Aspect-oriented programming (AOP) is a paradigm that promotes better modularity by providing means to encapsulate cross-cutting concerns in entities called aspects. Over the last years, a number of distributed aspect-oriented programming languages and systems have been proposed, illustrating the benefits of AOP in a distributed setting. Chemical calculi are particularly well-suited to formally specify the behaviour of concurrent and distributed systems. The join calculus is a functional name-passing calculus, with both distributed and object-oriented extensions. It is used as the basis of concurrency and distribution features in several mainstream languages like C# (Polyphonic C#, now Cω), OCaml (JoCaml), and Scala Joins. Unsurprisingly, practical programming in the join calculus also suffers from modularity issues when dealing with crosscutting concerns. We propose the Aspect Join Calculus [9], an aspect-oriented and distributed variant of the join calculus that addresses crosscutting and provides a formal foundation for distributed AOP. We develop a minimal aspect join calculus that allows aspects to advise chemical reactions. We show how to deal with causal relations in pointcuts and how to support advanced customisable aspect weaving semantics.

6.2. Type Theory and Proof Assistants

Participants: Simon Boulier, Gaëtan Gilbert, Maxime Lucas, Pierre-Marie Pédrot, Loïc Pujet, Nicolas Tabareau, Théo Winterhalter.

6.2.1. Type Theory

6.2.1.1. Effects in Type Theory.

There is a critical tension between substitution, dependent elimination and effects in type theory. In this paper, we crystallize this tension in the form of a no-go theorem that constitutes the fire triangle of type theory. To release this tension, we propose in [7] DCBPV, an extension of call-by-push-value (CBPV)-a general calculus of effects-to dependent types. Then, by extending to CBPV the well-known decompositions of call-by-name and call-by-value into CBPV, we show why, in presence of effects, dependent elimination must be restricted in call-by-name, and substitution must be restricted in call-by-value. To justify DCBPV and show that it is general enough to interpret many kinds of effects, we define various effectful syntactic translations from DCBPV to Martin-Löf type theory: the reader, weaning and forcing translations.

Traditional approaches to compensate for the lack of exceptions in type theories for proof assistants have severe drawbacks from both a programming and a reasoning perspective. We recently extended the Calculus of Inductive Constructions (CIC) with exceptions. The new exceptional type theory is interpreted by a translation into CIC, covering full dependent elimination, decidable type-checking and canonicity. However, the exceptional theory is inconsistent as a logical system. To recover consistency, we propose an additional translation that uses parametricity to enforce that all exceptions are caught locally. While this enforcement brings logical expressivity gains over CIC, it completely prevents reasoning about exceptional programs such as partial functions. In [6], we addresses the dilemma between exceptions and consistency in a more flexible manner, with the Reasonably Exceptional Type Theory (RETT). RETT is structured in three layers: (a) the exceptional layer, in which all terms can raise exceptions; (b) the mediation layer, in which exceptional terms must be provably parametric; (c) the pure layer, in which terms are non-exceptional, but can refer to exceptional terms. We present the general theory of RETT, where each layer is realized by a predicative hierarchy of universes, and develop an instance of RETT in Coq: the impure layer corresponds to the predicative universe hierarchy, the pure layer is realized by the impredicative universe of propositions, and the mediation layer is reified via a parametricity type class. RETT is the first full dependent type theory to support consistent reasoning about exceptional terms, and the CoqRETT plugin readily brings this ability to Coq programmers.

6.2.1.2. Eliminating Reflection from Type Theory.

Type theories with equality reflection, such as extensional type theory (ETT), are convenient theories in which to formalise mathematics, as they make it possible to consider provably equal terms as convertible. Although type-checking is undecidable in this context, variants of ETT have been implemented, for example in NuPRL and more recently in Andromeda. The actual objects that can be checked are not proof-terms, but derivations of proof-terms. This suggests that any derivation of ETT can be translated into a typecheckable proof term of intensional type theory (ITT). However, this result, investigated categorically by Hofmann in 1995, and 10 years later more syntactically by Oury, has never given rise to an effective translation. In [15], we provide the first syntactical translation from ETT to ITT with uniqueness of identity proofs and functional extensionality. This translation has been defined and proven correct in Coq and yields an executable plugin that translates a derivation in ETT into an actual Coq typing judgment. Additionally, we show how this result is extended in the context of homotopy to a two-level type theory.

6.2.1.3. Setoid type theory - a syntactic translation

[11] introduces setoid type theory, an intensional type theory with a proof-irrelevant universe of propositions and an equality type satisfying function extensionality, propositional extensionality and a definitional computation rule for transport. We justify the rules of setoid type theory by a syntactic translation into a pure type theory with a universe of propositions. We conjecture that our syntax is complete with regards to this translation.

6.2.1.4. The folk model category structure on strict ω-categories is monoidal

In [19], we prove that the folk model category structure on the category of strict ω-categories, introduced by Lafont, Métayer and Worytkiewicz, is monoidal, first, for the Gray tensor product and, second, for the join of ω-categories, introduced by the first author and Maltsiniotis. We moreover show that the Gray tensor

product induces, by adjunction, a tensor product of strict (m, n)-categories and that this tensor product is also compatible with the folk model category structure. In particular, we get a monoidal model category structure on the category of strict ω-groupoids. We prove that this monoidal model category structure satisfies the monoid axiom, so that the category of Gray monoids, studied by the second author, bears a natural model category structure.

6.2.2. Proof Assistants

6.2.2.1. Metacoq

The MetaCoq project [26], [26] aims to provide a certified meta-programming environment in Coq. It builds on Template-Coq, a plugin for Coq originally implemented by Malecha (2014), which provided a reifier for Coq terms and global declarations, as represented in the Coq kernel, as well as a denotation command. Recently, it was used in the CertiCoq certified compiler project (Anand et al., 2017), as its front-end language, to derive parametricity properties (Anand and Morrisett, 2018). However, the syntax lacked semantics, be it typing semantics or operational semantics, which should reflect, as formal specifications in Coq, the semantics of Coq's type theory itself. The tool was also rather bare bones, providing only rudimentary quoting and unquoting commands. We generalize it to handle the entire Polymorphic Calculus of Cumulative Inductive Constructions (pCUIC), as implemented by Coq, including the kernel's declaration structures for definitions and inductives, and implement a monad for general manipulation of Coq's logical environment. We demonstrate how this setup allows Coq users to define many kinds of general purpose plugins, whose correctness can be readily proved in the system itself, and that can be run efficiently after extraction. We give a few examples of implemented plugins, including a parametricity translation and a certifying extraction to call-by-value λ-calculus. We also advocate the use of MetaCoq as a foundation for higher-level tools.

6.2.2.2. Verification of Type Checking and Erasure for Coq, in Coq

Coq is built around a well-delimited kernel that perfoms typechecking for definitions in a variant of the Calculus of Inductive Constructions (CIC). Although the metatheory of CIC is very stable and reliable, the correctness of its implementation in Coq is less clear. Indeed, implementing an efficient type checker for CIC is a rather complex task, and many parts of the code rely on implicit invariants which can easily be broken by further evolution of the code. Therefore, on average, one critical bug has been found every year in Coq. [8] presents the first implementation of a type checker for the kernel of Coq (without the module system and template polymorphism), which is proven correct in Coq with respect to its formal specification and axiomatisation of part of its metatheory. Note that because of Gödel's incompleteness theorem, there is no hope to prove completely the correctness of the specification of Coq inside Coq (in particular strong normalisation or canonicity), but it is possible to prove the correctness of the implementation assuming the correctness of the specification, thus moving from a trusted code base (TCB) to a trusted theory base (TTB) paradigm. Our work is based on the MetaCoq project which provides metaprogramming facilities to work with terms and declarations at the level of this kernel. Our type checker is based on the specification of the typing relation of the Polymorphic, Cumulative Calculus of Inductive Constructions (pCUIC) at the basis of Coq and the verification of a relatively efficient and sound type-checker for it. In addition to the kernel implementation, an essential feature of Coq is the so-called extraction: the production of executable code in functional languages from Coq definitions. We present a verified version of this subtle type-and-proof erasure step, therefore enabling the verified extraction of a safe type-checker for Coq.

6.2.2.3. Definitional Proof-Irrelevance without K.

Definitional equality—or conversion—for a type theory with a decidable type checking is the simplest tool to prove that two objects are the same, letting the system decide just using computation. Therefore, the more things are equal by conversion, the simpler it is to use a language based on type theory. Proof-irrelevance, stating that any two proofs of the same proposition are equal, is a possible way to extend conversion to make a type theory more powerful. However, this new power comes at a price if we integrate it naively, either by making type checking undecidable or by realising new axioms—such as uniqueness of identity proofs (UIP)—that are incompatible with other extensions, such as univalence. In [3], taking inspiration from homotopy type theory, we propose a general way to extend a type theory with definitional proof irrelevance,

in a way that keeps type checking decidable and is compatible with univalence. We provide a new criterion to decide whether a proposition can be eliminated over a type (correcting and improving the so-called singleton elimination of Coq) by using techniques coming from recent development on dependent pattern matching without UIP. We show the generality of our approach by providing implementations for both Coq and Agda, both of which are planned to be integrated in future versions of those proof assistants.

6.2.2.4. Cubical Synthetic Homotopy Theory

Homotopy type theory is an extension of type theory that enables synthetic reasoning about spaces and homotopy theory. This has led to elegant computer formalizations of multiple classical results from homotopy theory. However, many proofs are still surprisingly complicated to formalize. One reason for this is the axiomatic treatment of univalence and higher inductive types which complicates synthetic reasoning as many intermediate steps, that could hold simply by computation, require explicit arguments. Cubical type theory offers a solution to this in the form of a new type theory with native support for both univalence and higher inductive types. In [14], we show how the recent cubical extension of Agda can be used to formalize some of the major results of homotopy type theory in a direct and elegant manner.

6.3. Program Certifications and Formalisation of Mathematics

Participants: Julien Cohen, Rémi Douence, Guilhem Jaber, Assia Mahboubi, Igor Zhirkov.

6.3.1. CoqTL: A Coq DSL for Rule-Based Model Transformation

In model-driven engineering, model transformation (MT) verification is essential for reliably producing software artifacts. While recent advancements have enabled automatic Hoare-style verification for non-trivial MTs, there are certain verification tasks (e.g. induction) that are intrinsically difficult to automate. Existing tools that aim at simplifying the interactive verification of MTs typically translate the MT specification (e.g. in ATL) and properties to prove (e.g. in OCL) into an interactive theorem prover. However, since the MT specification and proof phases happen in separate languages, the proof developer needs a detailed knowledge of the translation logic. Naturally, any error in the MT translation could cause unsound verification, i.e. the MT executed in the original environment may have different semantics from the verified MT. In [2] , we propose an alternative solution by designing and implementing an internal domain specific language, namely CoqTL, for the specification of declarative MTs directly in the Coq interactive theorem prover. Expressions in CoqTL are written in Gallina (the specification language of Coq), increasing the possibilities of reusing native Coq libraries in the transformation definition and proof. CoqTL specifications can be directly executed by our transformation engine encoded in Coq, or a certified implementation of the transformation can be generated by the native Coq extraction mechanism. We ensure that CoqTL has the same expressive power of Gallina (i.e. if a MT can be computed in Gallina, then it can also be represented in CoqTL). In this article, we introduce CoqTL, evaluate its practical applicability on a use case, and identify its current limitations.

6.3.2. A certificate-based approach to formally verified approximations.

In [12], we present a library to verify rigorous approximations of univariate functions on real numbers, with the Coq proof assistant. Based on interval arithmetic, this library also implements a technique of validation a posteriori based on the Banach fixed-point theorem. We illustrate this technique on the case of operations of division and square root. This library features a collection of abstract structures that organise the specification of rigorous approximations, and modularise the related proofs. Finally, we provide an implementation of verified Chebyshev approximations, and we discuss a few examples of computations.

6.3.3. Formally Verified Approximations of Definite Integrals.

Finding an elementary form for an antiderivative is often a difficult task, so numerical integration has become a common tool when it comes to making sense of a definite integral. Some of the numerical integration methods can even be made rigorous: not only do they compute an approximation of the integral value but they also bound its inaccuracy. Yet numerical integration is still missing from the toolbox when performing formal proofs in analysis. In [5], we present an efficient method for automatically computing and proving bounds

on some definite integrals inside the Coq formal system. Our approach is not based on traditional quadrature methods such as Newton-Cotes formulas. Instead, it relies on computing and evaluating antiderivatives of rigorous polynomial approximations, combined with an adaptive domain splitting. Our approach also handles improper integrals, provided that a factor of the integrand belongs to a catalog of identified integrable functions. This work has been integrated to the CoqInterval library.

6.3.4. Reasoning about exact memory transformations induced by refactorings in CompCert C

[18] reports on our work in extending CompCert memory model with a relation to model relocations. It preserves undefined values unlike similar relations defined in CompCert. This relation commutes with memory operations. Our main contributions are the relation itself and mechanically checked proofs of its commutation properties. We intend to use this extension to construct and verify a refactoring tool for programs written in C.

6.3.5. Automating Contextual Equivalence for Higher-Order Programs with References

In [4], we have proposed a framework to study contextual equivalence of programs written in a call-by-value functional language with local integer references. It reduces the problem of contextual equivalence to the problem of non-reachability in a transition system of memory configurations. This reduction is complete for recursion-free programs. Restricting to programs that do not allocate references inside the body of functions, we have encoded this non-reachability problem as a set of constrained Horn clause that can then be checked for satisfiability automatically. Restricting furthermore to a language with finite data-types, we also get a new decidability result for contextual equivalence at any type.

HYCOMES Project-Team

6. New Results

6.1. Mathematical Foundations of Physical Systems Modeling Languages

Participants: Albert Benveniste, Benoît Caillaud, Mathias Malandain.

Modern modeling languages for general physical systems, such as Modelica or Simscape, rely on Differential Algebraic Equations (DAE), i.e., constraints of the form $f(\dot{x}, x, u) = 0$. This facilitates modeling from first principles of the physics. This year we completed the development of the mathematical theory needed to sound, on solid mathematical bases, the design of compilers and tools for DAE based physical modeling languages.

Unlike Ordinary Differential Equations (ODE, of the form $\dot{x} = g(x, u)$), DAE exhibit subtle issues because of the notion of *differentiation index* and related *latent equations*—ODE are DAE of index zero for which no latent equation needs to be considered. Prior to generating execution code and calling solvers, the compilation of such languages requires a nontrivial *structural analysis* step that reduces the differentiation index to a level acceptable by DAE solvers.

Multimode DAE systems, having multiple modes with mode-dependent dynamics and state-dependent mode switching, are much harder to deal with. The main difficulty is the handling of the events of mode change. Unfortunately, the large literature devoted to the numerical analysis of DAEs does not cover the multimode case, typically saying nothing about mode changes. This lack of foundations causes numerous difficulties to the existing modeling tools. Some models are well handled, others are not, with no clear boundary between the two classes. Basically, no tool exists that performs a correct structural analysis taking multiple modes and mode changes into account.

In our work, we developed a comprehensive mathematical approach supporting compilation and code generation for this class of languages. Its core is the *structural analysis of multimode DAE systems,* taking both multiple modes and mode changes into account. As a byproduct of this structural analysis, we propose well sound criteria for accepting or rejecting models at compile time.

For our mathematical development, we rely on *nonstandard analysis,* which allows us to cast hybrid systems dynamics to discrete time dynamics with infinitesimal step size, thus providing a uniform framework for handling both continuous dynamics and mode change events.

A big comprehensive document has been written, which will be finalized and submitted next year.

6.2. Structural analysis of multimode DAE systems

Participants: Albert Benveniste, Benoît Caillaud, Khalil Ghorbal, Mathias Malandain.

The Hycomes team has obtained two results related to the structural analysis of multimode DAE systems.

6.2.1. Impulsive behavior of multimode DAE systems

A major difficulty with multimode DAE systems are the commutations from one mode to another one when the number of equations may change and variables may exhibit impulsive behavior, meaning that not only the trajectory of the system may be discontinuous, but moreover, some variables may be Dirac measures at the instant of mode changes. In [7] , we compare two radically different approaches to the structural analysis problem of mode changes. The first one is a classical approach, for a restricted class of DAE systems, for which the existence and uniqueness of an impulsive state jump is proved. The second approach is based on nonstandard analysis and is proved to generalize the former approach, to a larger class of multimode DAE systems. The most interesting feature of the latter approach is that it defines the state-jump as the standardization of the solution of a system of system of difference equations, in the framework of nonstandard analysis.

6.2.2. An implicit structural analysis method for multimode DAE systems

Modeling languages and tools based on Differential Algebraic Equations (DAE) bring several specific issues that do not exist with modeling languages based on Ordinary Differential Equations. The main problem is the determination of the differentiation index and latent equations. Prior to generating simulation code and calling solvers, the compilation of a model requires a structural analysis step, which reduces the differentiation index to a level acceptable by numerical solvers.

The Modelica language, among others, allows hybrid models with multiple modes, mode-dependent dynamics and state-dependent mode switching. These Multimode DAE (mDAE) systems are much harder to deal with. The main difficulties are (i) the combinatorial explosion of the number of modes, and (ii) the correct handling of mode switchings.

The focus of the paper [31] is on the first issue, namely: How can one perform a structural analysis of an mDAE in all possible modes, without enumerating these modes? A structural analysis algorithm for mDAE systems is presented, based on an implicit representation of the varying structure of an mDAE. It generalizes J. Pryce's Σ-method [56] to the multimode case and uses Binary Decision Diagrams (BDD) to represent the mode-dependent structure of an mDAE. The algorithm determines, as a function of the mode, the set of latent equations, the leading variables and the state vector. This is then used to compute a mode-dependent block-triangular decomposition of the system, that can be used to generate simulation code with a mode-dependent scheduling of the blocks of equations.

This method has been implemented in the IsamDAE software. This has allowed the Hycomes team to evaluate the performance and scalability of the method on several examples. In particular, it has been possible to perform the structural analysis of systems with more than 750 equations and 10^{23} modes.

6.3. Functional Decision Diagrams: A Unifying Data Structure For Binary Decision Diagrams

Participants: Joan Thibault, Khalil Ghorbal.

Zero-suppressed binary Decision Diagram (ZDD) is a notable alternative data structure of Reduced Ordered Binary Decision Diagram (ROBDD) that achieves a better size compression rate for Boolean functions that evaluate to zero almost everywhere. Deciding *a priori* which variant is more suitable to represent a given Boolean function is as hard as constructing the diagrams themselves. Moreover, converting a ZDD to a ROBDD (or vice versa) often has a prohibitive cost. This observation could be in fact stated about almost all existing BDD variants as it essentially stems from the non-compatibility of the reduction rules used to build such diagrams. Indeed, they are neither interchangeable nor composable. In [8], we investigate a novel functional framework, termed Lambda Decision Diagram (LDD), that ambitions to classify the already existing variants as implementations of special LDD models while suggesting, in a principled way, new models that exploit application-dependant properties to further reduce the diagram's size. We show how the reduction rules we use locally capture the global impact of each variable on the output of the entire function. Such knowledge suggests a variable ordering that sharply contrasts with the static fixed global ordering in the already existing variants as well as the dynamic reordering techniques commonly used.

PACAP Project-Team

7. New Results

7.1. Compilation and Optimization

Participants: Loïc Besnard, Caroline Collange, Byron Hawkins, Erven Rohou, Bahram Yarahmadi.

7.1.1. Optimization in the Presence of NVRAM

Participants: Erven Rohou, Bahram Yarahmadi.

A large and increasing number of Internet-of-Things devices are not equipped with batteries and harvest energy from their environment. Many of them cannot be physically accessed once they are deployed (embedded in civil engineering structures, sent in the atmosphere or deep in the oceans). When they run out of energy, they stop executing and wait until the energy level reaches a threshold. Programming such devices is challenging in terms of ensuring memory consistency and guaranteeing forward progress.

7.1.1.1. Checkpoint Placement based Worst-Case Energy Consumption

Previous work has proposed to insert checkpoints in the program so that execution can resume from well-defined locations. We propose to define these checkpoint locations based on worst-case energy consumption of code sections, with limited additional effort for programmers. As our method is based upon worst-case energy consumption, we can guarantee memory consistency and forward progress.

This work has been presented at the Compas 2019 conference.

7.1.1.2. Dynamic Adaptive Checkpoint Placement

Previous work has proposed to back-up the volatile states which are necessary for resuming the program execution after power failures. They either do it at compile time by placing checkpoints into the control flow of the program or at runtime by leveraging voltage monitoring facilities and interrupts, so that execution can resume from well-defined locations after power failures. We propose for the first time a dynamic checkpoint placement strategy which delays checkpoint placement and specialization to the runtime and takes decisions based on the past power failures and execution paths that are taken. We evaluate our work on a TI MSP430 device, with different types of benchmarks as well as different uninterrupted intervals, and we measure the execution time. We show that our work can outperform compiler-based state-of-the-art with memory footprint kept under the control.

This research is done within the context of the project IPL ZEP.

7.1.2. Dynamic Binary Optimization

Participant: Erven Rohou.

7.1.2.1. Guided just-in-time specialization

JavaScript's portability across a vast ecosystem of browsers makes it today a core building block of the web. Yet, building efficient systems in JavaScript is still challenging. Because this language is so dynamic, JavaScript programs provide little information that just-in-time compilers can use to carry out safe optimizations. Motivated by this observation, we propose to guide the JIT compiler in the task of code specialization. To this end, we have augmented [17] the language with an annotation that indicates which function call sites are likely to benefit from specialization. To support the automatic annotation of programs, we have introduced a novel static analysis that identifies profitable specialization points. We have implemented our ideas in JavaScriptCore, the built-in JavaScript engine for WebKit. The addition of guided specialization to this engine required us to change it in several non-trivial ways. Such changes let us observe speedups of up to 1.7× on programs present in synthetic benchmarks.

7.1.2.2. Run-time parallelization and de-parallelization

Runtime compilation has opportunities to parallelize code which are generally not available using static parallelization approaches. However, the parallelized code can possibly slowdown the performance due to unforeseen parallel overheads such as synchronization and speculation support pertaining to the chosen parallelization strategy and the underlying parallel platform. Moreover, with the wide usage of heterogeneous architectures, such choice options become more pronounced. We consider [22] an adaptive form of the parallelization operation, for the first time. We propose a method for performing on-stack de-parallelization for a parallelized binary loop at runtime, thereby allowing for rapid loop replacement with a more optimized one. We consider a loop parallelization strategy and propose a corresponding de-parallelization method. The method relies on stopping the execution at safe points, gathering threads' states, producing a corresponding serial code, and continuing execution serially. The decision to de-parallelize or not is taken based on the anticipated speedup. To assess the extent of our approach, we have conducted an initial study on a small set of programs with various parallelization overheads. Results show up to 4× performance improvement for a synchronization intense program on a 4-core Intel processor.

With the multicore trend, the need for automatic parallelization is more pronounced, especially for legacy and proprietary code where no source code is available and/or the code is already running and restarting is not an option. We engineer [21] a mechanism for transforming at runtime a frequent for-loop with no data dependencies in a binary program into a parallel loop, using on-stack replacement. With our mechanism, there is no need for source code, debugging information or restarting the program. Also, the mechanism needs no static instrumentation or information. The mechanism is implemented using the Padrone binary modification system and `pthreads`, where the remaining iterations of the loop are executed in parallel. The mechanism keeps the running program state by extracting the targeted loop into a separate function and copying the current stack frame into the corresponding frames of the created threads. Initial study is conducted on a set of kernels from the Polybench workload. Experimental results show from 2× to 3.5× speedup from sequential to parallelized code on four cores, which is similar to source code level parallelization.

This research was partially done within the context of the project PHC IMHOTEP.

7.1.3. Automatic and Parametrizable Memoization

Participants: Loïc Besnard, Erven Rohou.

Improving execution time and energy efficiency is needed for many applications and usually requires sophisticated code transformations and compiler optimizations. One of the optimization techniques is memoization, which saves the results of computations so that future computations with the same inputs can be avoided. We propose [16] a framework that automatically applies memoization techniques to C/C++ applications. The framework is based on automatic code transformations using a source-to-source compiler and on a memoization library. With the framework users can select functions to memoize as long as they obey to certain restrictions imposed by our current memoization library. We show the use of the framework and associated memoization technique and the impact on reducing the execution time and energy consumption of four representative benchmarks. The support library is available at https://gforge.inria.fr/projects/memoization (registered with APP under number IDDN.FR.001.250029.000.S.P.2018.000.10800).

7.1.4. Autotuning

Participants: Loïc Besnard, Erven Rohou.

The ANTAREX FET HPC project relies on a Domain Specific Language (DSL) based on Aspect Oriented Programming (AOP) concepts to allow applications to enforce extra functional properties such as energy-efficiency and performance and to optimize Quality of Service (QoS) in an adaptive way. The DSL approach allows the definition of energy-efficiency, performance, and adaptivity strategies as well as their enforcement at runtime through application autotuning and resource and power management. We present [20] an overview of the key outcome of the project, the ANTAREX DSL, and some of its capabilities through a number of examples, including how the DSL is applied in the context of the project use cases. We demonstrated [30] tools and techniques in two domains: computational drug discovery, and online vehicle navigation.

7.1.5. Loop splitting

The loop splitting technique takes advantage of long running loops to explore the impact of several optimization sequences at once, thus reducing the number of necessary runs. We rely on a variant of loop peeling which splits a loop into into several loops, with the same body, but a subset of the iteration space. New loops execute consecutive chunks of the original loop. We then apply different optimization sequences on each loop independently. Timers around each chunk observe the performance of each fragment. This technique may be generalized to combine compiler options and different implementations of a function called in a loop. It is useful when, for example, the profiling of the application shows that a function is critical in term of time of execution. In this case, the user must try to find the best implementation of their algorithm.

This research was partially done within the context of the ANTAREX FET HPC collaborative project, collaboration is currently ongoing with University of Porto, Portugal.

7.1.6. Hardware/Software JIT Compiler

Participant: Erven Rohou.

Single-ISA heterogeneous systems (such as ARM big.LITTLE) are an attractive solution for embedded platforms as they expose performance/energy trade-offs directly to the operating system. Recent works have demonstrated the ability to increase their efficiency by using VLIW cores, supported through Dynamic Binary Translation (DBT) to maintain the illusion of a single-ISA system. However, VLIW cores cannot rival with Out-of-Order (OoO) cores when it comes to performance, mainly because they do not use speculative execution. We study [27] how it is possible to use memory dependency speculation during the DBT process. Our approach enables fine-grained speculation optimizations thanks to a combination of hardware and software. Our results show that our approach leads to a geo-mean speed-up of 10 % at the price of a 7 % area overhead.

Our previous work on Hybrid-DBT was also presented at the RISC-V workshop in Zürich, Switzerland [38].

This work is a collaboration with the CAIRN team.

7.1.7. Scalable program tracing

Participants: Byron Hawkins, Erven Rohou.

The initial goal of scalable tracing is to record long executions at under $5\times$ overhead (ideally $2\times$), but it is equally important for analysis of the compressed trace to be efficient. This requires careful organization of the recorded data structures so that essential factors can be accessed without decompressing the trace or comprehensively iterating its paths. Precise context sensitivity is especially important for both optimization and security applications of trace-based program analysis, but scalability becomes challenging for frequently invoked functions that have a high degree of internal complexity. To avoid state space explosion in the context graph, such a function can be represented as a singleton while its complexity is preserved orthogonally. The current efforts focus mainly on developing an integration strategy to simplify program analysis over these two orthogonal dimensions of the trace.

7.1.8. Compiler optimization for quantum architectures

Participant: Caroline Collange.

In 2016, the first quantum processors have been made available to the general public. The possibility of programming an actual quantum device has elicited much enthusiasm [34]. Yet, such possibility also brought challenges. One challenge is the so called Qubit Allocation problem: the mapping of a virtual quantum circuit into an actual quantum architecture. There exist solutions to this problem; however, in our opinion, they fail to capitalize on decades of improvements on graph theory.

In collaboration with the Federal University of Minas Gerais, Brazil, we show how to model qubit allocation as the combination of Subgraph Isomorphism and Token Swapping [31]. This idea has been made possible by the publication of an approximative solution to the latter problem in 2016. We have compared our algorithm against five other qubit allocators, all independently designed in the last two years, including the winner of the IBM Challenge. When evaluated in "Tokyo", a quantum architecture with 20 qubits, our technique outperforms these state-of-the-art approaches in terms of the quality of the solutions that it finds and the amount of memory that it uses, while showing practical runtime.

7.2. Processor Architecture

Participants: Arthur Blanleuil, Niloofar Charmchi, Caroline Collange, Kleovoulos Kalaitzidis, Pierre Michaud, Anis Peysieux, Daniel Rodrigues Carvalho, André Seznec.

7.2.1. Value prediction

Participants: Kleovoulos Kalaitzidis, André Seznec.

Modern context-based value predictors tightly associate recurring values with instructions and contexts by building confidence upon them [9]. However, when execution monotony exists in the form of intervals, the potential prediction coverage is limited, since prediction confidence is reset at the beginning of each new interval. In [25], we address this challenge by introducing the notion of Equality Prediction (EP), which represents the binary facet of value prediction. Following a two fold decision scheme (similar to branch prediction), EP makes use of control-flow history to determine equality between the last committed result read at fetch time, and the result of the fetched occurrence. When equality is predicted with high confidence, the read value is used. Our experiments show that this technique obtains the same level of performance as previously proposed state-of-the-art context-based predictors. However, by virtue of better exploiting patterns of interval equality, our design complements the established way that value prediction is performed, and when combined with contemporary prediction models, improves the delivered speedup by 19 % on average.

7.2.2. Compressed caches

Participants: Daniel Rodrigues Carvalho, Niloofar Charmchi, Caroline Collange, André Seznec.

The speed gap between CPU and memory is impairing performance. Cache compression and hardware prefetching are two techniques that could confront this bottleneck by decreasing last level cache misses. However, compression and prefetching have positive interactions, as prefetching benefits from higher cache capacity and compression increases the effective cache size. We propose Compressed cache Layout Aware Prefetching (CLAP) to leverage the recently proposed sector-based compressed cache layouts such as SCC or YACC to create a synergy between compressed cache and prefetching. The idea of this approach is to prefetch contiguous blocks that can be compressed and co-allocated together with the requested block on a miss access [33]. Prefetched blocks that share storage with existing blocks do not need to evict a valid existing entry; therefore, CLAP avoids cache pollution. In order to decide the co-allocatable blocks to prefetch, we propose a compression predictor. Based on our experimental evaluations, CLAP reduces the number of cache misses by 12 % and improves performance by 4 % on average, comparing to a compressed cache [23].

7.2.3. Deep microarchitecture

Participants: Anis Peysieux, André Seznec.

The design of an efficient out-of-order execution core is particularly challenging. When the issue-width increases, the cost of the extra logic required by out-of-core execution increases dramatically. The silicon area occupied by this OoO core tends to grow quasi-quadratically with the issue-width (e.g. issue logic, register file and result bypass). At the same time, the power requirement and the energy consumption of the out-of–order core grow super-linearly with issue width. On wide-issue out-of-order execution cores, issue logic response time, register file access time, as well as result bypass delays represent potential critical paths that might impair cycle time or might necessitate further deepening of the execution pipeline. The objective of the PhD thesis of Anis Peysieux will be to reduce the number of instructions that enter the OoO core, and therefore to master the hardware complexity while still achieving the performance promises of a very wide issue processor.

7.2.4. Dynamic thermal management

Participant: Pierre Michaud.

As power dissipation and circuit temperature constrain their performance, modern processors feature turbo control mechanisms to adjust the voltage and clock frequency dynamically so that circuit temperature stays below a certain limit. In particular, turbo control exploits the fact that, after a long period of low processor activity, the thermal capacity of the chip, its package and the heatsink can absorb heat at a relatively fast rate during a certain time, before the temperature limit constrains that rate. Hence power dissipation can be temporarily boosted above the average sustainable value. The turbo control must monitor circuit temperature continuously to maximize the clock frequency. Temperature can be monitored by reading the integrated thermal sensors. However, making the clock frequency depend on thermal sensor readings implies that processor performance depends on ambient temperature. Yet this form of performance non-determinism is a problem for certain processor makers. A possible solution is to determine the clock frequency not from the true temperature but from a thermal model based on the nominal ambient temperature. Such model should be as accurate as possible in order to prevent sensor-based protection from triggering but sporadically, without hurting performance by overestimating temperature too much. The model should also be simple enough to provide calculated temperature in real time. We propose a thermal model possessing these qualities, and a new turbo control algorithm based on that model [37].

7.2.5. Thread convergence prediction for general-purpose SIMT architectures

Participants: Arthur Blanleuil, Caroline Collange.

GPUs group threads of SPMD programs in warps and synchronize them to execute the same instruction at the same time. This execution model, referred to as Single-Instruction, Multiple-Thread (SIMT), enables the use of energy-efficient SIMD execution units by factoring out control logic such as instruction fetch and decode pipeline stages for a whole warp. SIMT execution is the key enabler for the energy efficiency of GPUs. We seek to generalize the SIMT execution model to general-purpose superscalar cores.

As threads within a warp may follow different directions through conditional branches in the program, the warp must follow each taken path in turn, while disabling individual threads that do not participate. Following divergence, current GPU architectures attempt to restore convergence at the earliest program point following static annotations in the binary. However, this policy has been shown to be suboptimal in many cases, in which later convergence improves performance. In fact, optimal convergence points depend on dynamic program behavior, so static decisions are unable to capture them.

The goal of the thesis of Arthur Blanleuil is to design predictors that enable the microarchitecture to infer dynamic code behavior and place convergence points appropriately. Convergence predictors have analogies with branch predictors and control independence predictors studied in superscalar processor architecture, but they present one additional challenge: the thread runaway problem. Although a branch misprediction will be identified and repaired locally, a wrong thread scheduling decision may go unnoticed and delay convergence by thousands of instructions. To address the thread runaway problem, we plan to explore promise-based speculation and recovery strategies. When no information is available, we follow the traditional conservative earliest-convergence scheduling policy. Once the predictor has enough information to make a more aggressive prediction, it generates assumptions about the prediction. The microarchitecture then keeps checking dynamically whether the assumptions actually hold true in the near future. If assumptions turn out to be wrong, the prediction will be reconsidered by changing back priorities to conservative. Such promise-based speculation policies can address the thread runaway problem by fixing a bound on the worst-case performance degradation of an aggressive scheduling policy against the conservative baseline.

Accurate thread convergence policies will enable dynamic vectorization to adapt to application characteristics dynamically. They will both improve performance and simplify programming of many-core architectures by alleviating the need for advanced code tuning by expert programmers.

7.2.6. Exploring the design space of GPU architectures

Participants: Alexandre Kouyoumdjian, Caroline Collange.

We study tradeoffs in the internal organization of GPUs in the context of general-purpose parallel processing [35]. In particular, we analyze the performance impact of having a few wide streaming multiprocessors compared to many narrow ones. Although we find narrow configurations usually give higher performance for an equal number of execution units, they require more hardware resources and energy. On the other hand, our evaluation show that the optimal streaming multiprocessor width varies across applications. This study motivates adaptive GPU architectures that would support configurable internal organization.

7.3. WCET estimation and optimization

Participants: Loïc Besnard, Damien Hardy, Isabelle Puaut, Stefanos Skalistis.

7.3.1. WCET estimation for many core processors

Participants: Damien Hardy, Isabelle Puaut, Stefanos Skalistis.

7.3.1.1. Optimization of WCETs by considering the effects of local caches

The overall goal of this research is to define WCET estimation methods for parallel applications running on many-core architectures, such as the Kalray MPPA machine. Some approaches to reach this goal have been proposed, but they assume the mapping of parallel applications on cores is already done. Unfortunately, on architectures with caches, task mapping requires a priori known WCETs for tasks, which in turn requires knowing task mapping (i.e., co-located tasks, co-running tasks) to have tight WCET bounds. Therefore, scheduling parallel applications and estimating their WCET introduce a chicken-and-egg situation.

We addressed this issue by developing both optimal and heuristic techniques for solving the scheduling problem, whose objective is to minimize the WCET of a parallel application. Our proposed static partitioned non-preemptive mapping strategies address the effect of local caches to tighten the estimated WCET of the parallel application. Experimental results obtained on real and synthetic parallel applications show that co-locating tasks that reuse code and data improves the WCET by 11 % on average for the optimal method and by 9 % on average for the heuristic method. An implementation on the Kalray MPPA machine allowed to identify implementation-related overheads. All results are described in [18].

7.3.1.2. Shared resource contentions and WCET estimation

Accurate WCET analysis for multi-cores is known to be challenging, because of concurrent accesses to shared resources, such as communication through busses or Networks on Chips (NoC). Since it is impossible in general to guarantee the absence of resource conflicts during execution, current WCET techniques either produce pessimistic WCET estimates or constrain the execution to enforce the absence of conflicts, at the price of a significant hardware under-utilization. In addition, the large majority of existing works consider that the platform workload consists of independent tasks. As parallel programming is the most promising solution to improve performance, we envision that within only a few years from now, real-time workloads will evolve toward parallel programs. The WCET behavior of such programs is challenging to analyze because they consist of *dependent* tasks interacting through complex synchronization/communication mechanisms.

In [28], we propose a scheduling technique that jointly selects Scratchpad Memory (SPM) contents off-line, in such a way that the cost of SPM loading/unloading is hidden. Communications are fragmented to augment hiding possibilities. Experimental results show the effectiveness of the proposed technique on streaming applications and synthetic task-graphs. The overlapping of communications with computations allows the length of generated schedules to be reduced by 4 % on average on streaming applications, with a maximum of 16 %, and by 8 % on average for synthetic task graphs. We further show on a case study that generated schedules can be implemented with low overhead on a predictable multi-core architecture (Kalray MPPA).

7.3.1.3. Interference-sensitive run-time adaptation of time-triggered schedules

In time-critical systems, run-time adaptation is required to improve the performance of time-triggered execution, derived based on Worst-Case Execution Time (WCET) of tasks. By improving performance, the systems can provide higher Quality-of-Service, in safety-critical systems, or execute other best-effort applications, in mixed-critical systems. To achieve this goal, we propose in [32] a parallel interference-sensitive run-time adaptation mechanism that enables a fine-grained synchronisation among cores. Since the run-time adaptation of

offline solutions can potentially violate the timing guarantees, we present the Response-Time Analysis (RTA) of the proposed mechanism showing that the system execution is free of timing-anomalies. The RTA takes into account the timing behavior of the proposed mechanism and its associated WCET. To support our contribution, we evaluate the behavior and the scalability of the proposed approach for different application types and execution configurations on the 8-core Texas Instruments TMS320C6678 platform. The obtained results show significant performance improvement compared to state-of-the-art centralized approaches.

7.3.1.4. WCET-Aware Parallelization of Model-Based Applications for Multi-Cores

Parallel architectures are nowadays not only confined to the domain of high performance computing, they are also increasingly used in embedded time-critical systems.

The Argo H2020 project provides a programming paradigm and associated tool flow to exploit the full potential of architectures in terms of development productivity, time-to-market, exploitation of the platform computing power and guaranteed real-time performance. The Argo toolchain operates on Scilab and XCoS inputs, and targets ScratchPad Memory (SPM)-based multi-cores. Data-layout and loop transformations play a key role in this flow as they improve SPM efficiency and reduce the number of accesses to shared main memory.

In [19] we present the overall results of the project, a compiler tool-flow for automated parallelization of model-based real-time software, which addresses the shortcomings of multi-core architectures in real-time systems. The flow is demonstrated using a model-based Terrain Awareness and Warning Systems (TAWS) and an edge detection algorithm from the image-processing domain. Model-based applications are first transformed into real-time C code and from there into a well-predictable parallel C program. Tight bounds for the Worst-Case Execution Time (WCET) of the parallelized program can be determined using an integrated multi-core WCET analysis. Thanks to the use of an architecture description language, the general approach is applicable to a wider range of target platforms. An experimental evaluation for a research architecture with network-on-chip (NoC) interconnect shows that the parallel WCET of the TAWS application can be improved by factor 1.77 using the presented compiler tools.

7.3.2. WCET estimation and optimizing compilers

Participants: Isabelle Puaut, Stefanos Skalistis.

Static Worst-Case Execution Time (WCET) estimation techniques operate upon the binary code of a program in order to provide the necessary input for schedulability analysis techniques. Compilers used to generate this binary code include tens of optimizations, that can radically change the flow information of the program. Such information is hard to be maintained across optimization passes and may render automatic extraction of important flow information, such as loop bounds, impossible. Thus, compiler optimizations, especially the sophisticated optimizations of mainstream compilers, are typically avoided. We explore [24] for the first time iterative-compilation techniques that reconcile compiler optimizations and static WCET estimation. We propose a novel learning technique that selects sequences of optimizations that minimize the WCET estimate of a given program. We experimentally evaluate the proposed technique using an industrial WCET estimation tool (AbsInt aiT) over a set of 46 benchmarks from four different benchmarks suites, including reference WCET benchmark applications, image processing kernels and telecommunication applications. Experimental results show that WCET estimates are reduced on average by 20.3 % using the proposed technique, as compared to the best compiler optimization level applicable.

7.3.3. WCET estimation and processor micro-architecture

Participant: Isabelle Puaut.

Cache memories in modern embedded processors are known to improve average memory access performance. Unfortunately, they are also known to represent a major source of unpredictability for hard real-time workload. One of the main limitations of typical caches is that content selection and replacement is entirely performed in hardware. As such, it is hard to control the cache behavior in software to favor caching of blocks that are known to have an impact on an application's worst-case execution time (WCET). In [26], we consider a cache replacement policy, namely DM-LRU, that allows system designers to prioritize caching of memory blocks that are known to have an important impact on an application's WCET. Considering a single-core, single-level

cache hierarchy, we describe an abstract interpretation-based timing analysis for DM-LRU. We implement the proposed analysis in a self-contained toolkit and study its qualitative properties on a set of representative benchmarks. Apart from being useful to compute the WCET when DM-LRU or similar policies are used, the proposed analysis can allow designers to perform WCET impact-aware selection of content to be retained in cache.

Long pipelines need good branch predictors to keep the pipeline running. Current branch predictors are optimized for the average case, which might not be a good fit for real-time systems and worst-case execution time analysis. We present [29] a time-predictable branch predictor co-designed with the associated worst-case execution time analysis. Thee branch predictor uses a fully-associative cache to track branch outcomes and destination addresses. The fully-associative cache avoids any false sharing of entries between branches. Therefore, we can analyze program scopes that contain a number of branches lower than or equal to the number of branches in the prediction table. Experimental results show that the worst-case execution time bounds of programs using the proposed predictor are lower than using static branch predictors at a moderate hardware cost.

7.4. Security

Participants: Nicolas Bellec, Damien Hardy, Kévin Le Bon, Isabelle Puaut, Erven Rohou.

7.4.1. Attack detection co-processor for real-time systems

Participants: Nicolas Bellec, Isabelle Puaut.

Real-time embedded systems (RTES) are required to interact more and more with their environment, thereby increasing their attack surface. Recent security breaches on car brakes and other critical components, have already proven the feasibility of attacks on RTES. Such attacks may change the control-flow of the programs, which may lead to violations of the timing constraints of the system. In this ongoing work, we design a technique to detect attacks in RTES based on timing information. Our technique is based on a monitor, implemented in hardware to preserve the predictability of instrumented programs. The monitor uses timing information (Worst-Case Execution Time – WCET – of code regions) to detect attacks. An algorithm for the region selection, optimal when the monitoring memory is not limited is presented and provides guarantees on attack detection latency. An implementation of the hardware monitor and its simulation demonstrates the practicality of our approach. An experimental study evaluates the maximum attack detection latency for different monitor memory budgets.

This work is done in collaboration with the CIDRE and CAIRN teams.

7.4.2. Multi-nop fault injection attack

Participants: Damien Hardy, Erven Rohou.

The CIDRE team has developed a platform named Traitor that allows to perform multiple fault injection attack by replacing instructions by nops during the execution of a program. In this context, we are defining a program model where each instruction can be replaced by a nop at runtime. On this model we plan to apply compilation techniques on the binary to automatically determine where nops have to be inserted at runtime to perform sophisticated attacks such as dump of memory, modification of the memory, memory protection deactivation, execution of code in RAM.

This work is done in collaboration with the CIDRE team.

7.4.3. Compiler-based automation of side-channel countermeasures

Participants: Damien Hardy, Erven Rohou.

Masking is a popular protection against side-channel analysis exploiting the power consumption or electromagnetic radiations. Besides the many schemes based on simple Boolean encoding, some alternative schemes such as Orthogonal Direct Sum Masking (ODSM) or Inner Product Masking (IP) aim to provide more security, reduce the entropy or combine masking with fault detection. The practical implementation of those schemes is done manually at assembly or source-code level, some of them even stay purely theoretical. We proposed a compiler extension to automatically apply different masking schemes for block cipher algorithms. We introduced a generic approach to describe the schemes and we inserted three of them at compile-time on an AES implementation. Currently, a practical side-channel analysis is performed in collaboration with TAMIS to assess the correctness and the performance of the code inserted.

This work is done in collaboration with the TAMIS team.

7.4.4. Platform for adaptive dynamic protection of programs

Participants: Kévin Le Bon, Erven Rohou.

Memory corruption attacks are a serious threat for system integrity. Many techniques have been developed in order to protect systems from these attacks. However, the deployment of heavy protections often degrades the performance of programs. We propose [36] a dynamic approach that adapts the protection level of the target process during its execution depending on the observed behavior.

SUMO Project-Team

7. New Results

7.1. New results on Axis 1: Quantitative models

7.1.1. Verification of Real-Time Models

Participants : Ocan Sankur, Nicolas Markey, Victor Roussanaly

7.1.1.1. Abstraction-refinement algorithms for model checking of timed automata.

The abstraction domain we consider [26] abstracts away zones by restricting the set of clock constraints that can be used to define them, while the refinement procedure computes the set of constraints that must be taken into consideration in the abstraction so as to exclude a given spurious counterexample. We implement this idea in two ways: an enumerative algorithm where a lazy abstraction approach is adopted, meaning that possibly different abstract domains are assigned to each exploration node; and a symbolic algorithm where the abstract transition system is encoded with Boolean formulas.

7.1.1.2. Robust controller synthesis problem in Büchi timed automata

We solve a robust controller synthesis problem [20] in a purely symbolic way. The goal of the controller is to play according to an accepting lasso of the automaton, while resisting to timing perturbations chosen by a competing environment. The problem was previously shown to be *PSPACE*-complete using regions-based techniques, but we provide a first tool solving the problem using zones only, thus more resilient to state-space explosion problem. The key ingredient is the introduction of branching constraint graphs allowing to decide in polynomial time whether a given lasso is robust, and even compute the largest admissible perturbation if it is. We also make an original use of constraint graphs in this context in order to test the inclusion of timed reachability relations, crucial for the termination criterion of our algorithm. Our techniques are illustrated using a case study on the regulation of a train network.

7.1.2. Verification of Stochastic Models

Participants : Hugo Bazille, Nathalie Bertrand, Éric Fabre, Blaise Genest, Ocan Sankur

7.1.2.1. Long-run satisfaction of path properties

We introduced the concepts of long-run frequency of path properties for paths in Kripke structures, and their generalization to long-run probabilities for schedulers in Markov decision processes [13]. We then studied the natural optimization problem of computing the optimal values of these measures, when ranging over all paths or all schedulers, and the corresponding decision problem when given a threshold. The main results are as follows. For (repeated) reachability and other simple properties, optimal long-run probabilities and corresponding optimal memoryless schedulers are computable in polynomial time. When it comes to constrained reachability properties, memoryless schedulers are no longer sufficient, even in the non-probabilistic setting. Nevertheless, optimal long-run probabilities for constrained reachability are computable in pseudo-polynomial time in the probabilistic setting and in polynomial time for Kripke structures. Finally for co-safety properties expressed by NFA, we gave an exponential-time algorithm to compute the optimal long-run frequency, and proved the PSPACE-completeness of the threshold problem.

7.1.2.2. Approximate Verification of Dynamic Bayesian Networks.

We are interested in studying the evolution of large homogeneous populations of cells, where each cell is assumed to be composed of a group of biological players (species) whose dynamics is governed by a complex biological pathway, identical for all cells. Modeling the inherent variability of the species concentrations in different cells is crucial to understand the dynamics of the population. In [9], we focus on handling this variability by modeling each species by a random variable that evolves over time. This appealing approach runs into the curse of dimensionality since exactly representing a joint probability distribution involving a large set of random variables quickly becomes intractable as the number of variables grows. To make this approach amenable to biopathways, we explore different techniques to (i) approximate the exact joint distribution at a given time point, and (ii) to track its evolution as time elapses.

7.1.2.3. Classification among stochastic systems

An important task in AI is one of classifying an observation as belonging to one class among several (e.g. image classification). We revisit this problem in a verification context: given k partially observable systems modeled as Hidden Markov Models (HMMs, also called labeled Markov chains), and an execution of one of them, can we eventually classify which system performed this execution, just by looking at its observations? Interestingly, this problem generalizes several problems in verification and control, such as fault diagnosis and opacity. Also, classification has strong connections with different notions of distances between stochastic models.

In [12], we study a general and practical notion of classifiers, namely limit-sure classifiers, which allow misclassification, i.e. errors in classification, as long as the probability of misclassification tends to 0 as the length of the observation grows. To study the complexity of several notions of classification, we develop techniques based on a simple but powerful notion of stationary distributions for HMMs. We prove that one cannot classify among HMMs iff there is a finite separating word from their stationary distributions. This provides a direct proof that classifiability can be checked in PTIME, as an alternative to existing proofs using separating events (i.e. sets of infinite separating words) for the total variation distance. Our approach also allows us to introduce and tackle new notions of classifiability which are applicable in a security context.

7.1.2.4. Fault diagnosis for stochastic systems

Diagnosis of partially observable stochastic systems prone to faults was introduced in the late nineties. Diagnosability, *i.e.* the existence of a diagnoser, may be specified in different ways: exact diagnosability requires that almost surely a fault is detected and that no fault is erroneously claimed; approximate diagnosability tolerates a small error probability when claiming a fault; last, accurate approximate diagnosability guarantees that the error probability can be chosen arbitrarily small.

In the article [7], we first refine the specification of diagnosability by identifying three criteria: (1) detecting faulty runs or providing information for all runs (2) considering finite or infinite runs, and (3) requiring or not a uniform detection delay. We then give a complete picture of relations between the different diagnosability specifications for probabilistic systems and establish characterisations for most of them in the finite-state case. Based on these characterisations, we develop decision procedures, study their complexity and prove their optimality. We also design synthesis algorithms to construct diagnosers and we analyse their memory requirements. Finally we establish undecidability of the diagnosability problems for which we provided no characterisation.

7.1.3. Energy Games

Participants : Loïc Hélouët, Nicolas Markey

7.1.3.1. Games with reachability objectives under energy constraints.

Under strict energy constraints (either only lower-bound constraint or interval constraint), we prove [23] that games with reachability objectives are LOGSPACE-equivalent to energy games with the same energy constraints but without reachability objective (i.e., for infinite runs). We then consider two kinds of relaxations of the upper-bound constraints (while keeping the lower-bound constraint strict): in the first one, called weak upper bound, the upper bound is absorbing, in the sense that it allows receiving more energy when the upper bound is already reached, but the extra energy will not be stored; in the second one, we allow for temporary violations of the upper bound, imposing limits on the number or on the amount of violations. We prove that when considering weak upper bound, reachability objectives require memory, but can still be solved in polynomial-time for one-player arenas; we prove that they are in co-NP in the two-player setting. Allowing for bounded violations makes the problem PSPACE-complete for one-player arenas and EXPTIME-complete for two players.

7.2. New results on Axis 2: Large Systems Models

7.2.1. Smart Transportation Systems

Participants : Nathalie Bertrand, Loïc Hélouët, Ocan Sankur

7.2.1.1. Smart regulation of urban train systems.

We have considered application of model checking techniques to evaluate performances of urban train systems [15]. Metros are subject to unexpected delays due to weather conditions, incidents, passenger misconduct, etc. To recover from delays and avoid their propagation to the whole network, metro operators use regulation algorithms that adapt speeds and departure dates of trains. Regulation algorithms are ad-hoc tools tuned to cope with characteristics of tracks, rolling stock, and passengers habits. However, there is no universal optimal regulation adapted in any environment. So, performance of a regulation must be evaluated before its integration in a network. In this work, we use probabilistic model-checking to evaluate the performance of regulation algorithms in simple metro lines. We model the moves of trains and random delays with Markov decision processes, and regulation as a controller that forces a decision depending on its partial knowledge of the state of the system. We then use the probabilistic model checker PRISM to evaluate performance of regulation: We compute the probability to reach a stable situation from an unstable one in less than d time units, letting d vary in a large enough time interval. This approach is applied on a case study, the metro network of Glasgow.

7.2.2. Supervisory Control

Participants : Hervé Marchand

7.2.2.1. Towards resilient supervisors against sensor deception attacks.

As a security problem, we considered in [24] feedback control systems where sensor readings may be compromised by a malicious attacker intent on causing damage to the system. We study this problem at the supervisory layer of the control system, using discrete event systems techniques. We assume that the attacker can edit the outputs from the sensors of the system before they reach the supervisory controller. In this context, we formulate the problem of synthesizing a supervisor that is robust against a large class of edit attacks on the sensor readings. The solution methodology is based on the solution of a partially observed supervisory control problem with arbitrary control patterns.

7.2.3. Multi-agent systems

Participants : Arthur Queffelec, Nicolas Markey, Ocan Sankur

7.2.3.1. Multi-agent path planning problems.

We are motivated by the increasing appeal of robots in information-gathering missions. In the problems we study [21], [22], the agents must remain interconnected. We model an area by a topological graph specifying the movement and the connectivity constraints of the agents. We study the theoretical complexity of the reachability and the coverage problems of a fleet of connected agents on various classes of topological graphs. We establish the complexity of these problems on known classes, and introduce a new class called sight-moveable graphs which admit efficient algorithms.

7.2.3.2. Quantitative semantics for Strategy Logic

We introduce and study SL[F], a quantitative extension of SL (Strategy Logic) [19], one of the most natural and expressive logics describing strategic behaviours. The satisfaction value of an SL[F] formula is a real value in [0,1], reflecting "how much" or "how well" the strategic on-going objectives of the underlying agents are satisfied. We demonstrate the applications of SL[F] in quantitative reasoning about multi-agent systems, by showing how it can express concepts of stability in multi-agent systems, and how it generalises some fuzzy temporal logics. We also provide a model-checking algorithm for our logic, based on a quantitative extension of Quantified CTL.

7.3. New results on Axis 3: Population Models

7.3.1. Verification

Participants : Nathalie Bertrand, Anirban Majumdar

7.3.1.1. Networks of many identical agents communicating by broadcast.

Broadcast networks allow one to model networks of identical nodes communicating through message broadcasts [17]. Their parameterized verification aims at proving a property holds for any number of nodes, under any communication topology, and on all possible executions. We focus on the coverability problem which dually asks whether there exists an execution that visits a configuration exhibiting some given state of the broadcast protocol. Coverability is known to be undecidable for static networks, i.e. when the number of nodes and communication topology is fixed along executions. In contrast, it is decidable in PTIME when the communication topology may change arbitrarily along executions, that is for reconfigurable networks. Surprisingly, no lower nor upper bounds on the minimal number of nodes, or the minimal length of covering execution in reconfigurable networks, appear in the literature. We showed tight bounds for cutoff and length, which happen to be linear and quadratic, respectively, in the number of states of the protocol. We also introduced an intermediary model with static communication topology and non-deterministic message losses upon sending. We showed that the same tight bounds apply to lossy networks, although, reconfigurable executions may be linearly more succinct than lossy executions. Finally, we showed NP-completeness for the natural optimisation problem associated with the cutoff.

7.3.1.2. Randomized distributed algorithms for consensus.

Randomized fault-tolerant distributed algorithms pose a number of challenges for automated verification: (i) parameterization in the number of processes and faults, (ii) randomized choices and probabilistic properties, and (iii) an unbounded number of asynchronous rounds. This combination makes verification hard. Challenge (i) was recently addressed in the framework of threshold automata. We extended threshold automata to model randomized consensus algorithms that perform an unbounded number of asynchronous rounds. For non-probabilistic properties, we showed [18] that it is necessary and sufficient to verify these properties under round-rigid schedules, that is, schedules where processes enter round r only after all processes finished round $r-1$. For almost-sure termination, we analyzed these algorithms under round-rigid adversaries, that is, fair adversaries that only generate round-rigid schedules. This allowed us to do compositional and inductive reasoning that reduces verification of the asynchronous multi-round algorithms to model checking of a one-round threshold automaton. We applied this framework and automatically verified the following classic algorithms: Ben-Or's and Bracha's seminal consensus algorithms for crashes and Byzantine faults, 2-set agreement for crash faults, and RS-Bosco for the Byzantine case.

7.3.2. Control

Participants : Nathalie Bertrand, Blaise Genest, Anirban Majumdar

7.3.2.1. Controlling a population

We introduced a new setting where a population of agents [6], each modelled by a finite-state system, are controlled uniformly: the controller applies the same action to every agent. The framework is largely inspired by the control of a biological system, namely a population of yeasts, where the controller may only change the environment common to all cells. We studied a synchronisation problem for such populations: no matter how individual agents react to the actions of the controller, the controller aims at driving all agents synchronously to a target state. The agents are naturally represented by a non-deterministic finite state automaton (NFA), the same for every agent, and the whole system is encoded as a 2-player game. The first player (Controller) chooses actions, and the second player (Agents) resolves non-determinism for each agent. The game with m agents is called the m-population game. This gives rise to a parameterized control problem (where control refers to 2 player games), namely the population control problem: can Controller control the m-population game for all $m \in \mathbb{N}$ whatever Agents does? In this work, we proved that the population control problem is decidable, and it is a EXPTIME-complete problem. As far as we know, this is one of the first results on the control of parameterized systems. Our algorithm, which is not based on cut-off techniques, produces winning strategies which are symbolic, that is, they do not need to count precisely how the population is spread between states. The winning strategies produced by our algorithm are optimal with respect to the synchronisation time: the maximal number of steps before synchronisation of all agents in the target state is at most polynomial in the number of agents m, and exponential in the size of the NFA. We also showed that if there is no winning

strategy, then there is a population size M such that Controller wins the m-population game if and only if m $\leq$ M. Surprisingly, M can be doubly exponential in the number of states of the NFA, with tight upper and lower bounds.

7.3.2.2. Concurrent multiplayer games with arbitrary many players

Traditional concurrent games on graphs involve a fixed number of players, who take decisions simultaneously, determining the next state of the game. In [16], we introduced a parameterized variant of concurrent games on graphs, where the parameter is precisely the number of players. Parameterized concurrent games are described by finite graphs, in which the transitions bear regular languages to describe the possible move combinations that lead from one vertex to another. We considered the problem of determining whether the first player, say Eve, has a strategy to ensure a reachability objective against any strategy profile of her opponents as a coalition. In particular Eve's strategy should be independent of the number of opponents she actually has. Technically, we focused on an *a priori* simpler setting where the languages labeling transitions only constrain the number of opponents (but not their precise action choices). These constraints are described as semilinear sets, finite unions of intervals, or intervals. We established the precise complexities of the parameterized reachability game problem, ranging from PTIME-complete to PSPACE-complete, in a variety of situations depending on the contraints (semilinear predicates, unions of intervals, or intervals) and on the presence or not of non-determinism.

7.4. New results on Axis 4: Data-driven Models

7.4.1. Crowdsourcing

Participants : Loïc Hélouët, Rituraj Singh

7.4.1.1. Complex workflows for crowdsourcing.

Crowdsourcing consists in hiring workers on internet to perform large amounts of simple, independent and replicated work units. We have proposed [32] complex workflows, a model for concurrent orcestration of tasks to solve problems that are more intricate than simpe tagging problems. Complex workflows allow higher-order answers where workers can suggest a process to obtain data rather than a plain answer. It is a data-centric model based on orchestration of concurrent tasks and higher order schemes. We have considered formal properties of specifications described with this model termination (whether some/all runs of a complex workflow terminate) and correctness (whether some/all runs of a workflow terminate with data satisfying FO requirements). We have shown that existential termination/correctness are undecidable in general excepted for specifications with bounded recursion. However, universal termination/correctness are decidable when constraints on inputs are specified in a decidable fragment of FO, and are at least in 2EXPTIME.

7.4.1.2. CrowdInc : a solution to reduce the cost of Consensus in Crowdsourcing.

Another contribution around crowdsourcing [34] considers agregation of answers, reliability of computed results, and optimization of costs. Crowdsoucing call for human expertise to solve problems which are still hard for computers, but easy for human workers. Crowdsourcing platform distribute replicated tasks to workers, pay them for their contribution, and aggregate answers to produce a reliable conclusion. A fundamental problem is to infer a correct answer from the set of results returned by workers. An additional ingredient of crowdsourcing is the cost needed to obtain a reliable answer: unlimited budget allows for the use of large pools of workers for each task, or experts to improve reliability of aggregated answers, but a limited budget forces to use resources at best to synthesize an reasonably reliable answer. We have focused on crowdsourcing of simple tasks with boolean answers. In this setting, we have first defined a probabilistic inference technique to agregate answers. This allows to consider difficulty of tasks and expertise of workers when building a conclusion. We have then proposed a greedy algorithm that reduces the cost (i.e. the number of workers hired by a platform) needed to reach a consensual answer. This algorithm considers difficulty of task, budget provided by client and total tasks to dynamically adapt threshold at each stage and makes locally optimal choice while preserving accuracy. Last, we have shown efficiency of our algorithm on several benchmarks, and compared its performance to existing solutions.

7.4.2. Guarded Attribute Grammars and Petri net synthesis

Participants : Adrian Puerto Aubel, Éric Badouel

7.4.2.1. Service-oriented programming

We addressed [30] the problem of component reuse for the design of user-centric distributed collaborative systems modelled by Guarded Attribute Grammars. Following the contract-based specification of components we develop an approach to an interface theory for the components of a collaborative system in three stages: we define a composition of interfaces that specifies how the component behaves with respect to its environement, we introduce an implementation order on interfaces and finally a residual operation on interfaces characterizing the systems that, when composed with a given component, can complement it in order to realize a global specification.

The visit of Joskel Ngoufo, a doctoral student at Yaoundé University, was the occasion to initiate a new implementation of the Guarded Attribute Grammars engine, in Racket language, a dialect of Lisp that allows metalanguage facilities and graphical interfaces to be processed more easily than in Haskell, the language chosen for the previous implementation.

7.4.2.2. Coordination of public debate.

Our research on data-centric collaborative systems has focused this year on the modelling of debates [28], with the aim of producing a tool that makes it possible to automatically conduct them, while managing relevant documents and analysing the respective positions of the different interventions from the point of view of argumentation theory. To this end, we are collaborating with Carlo Ferigato, a researcher at the JRC (C.E. Ispra, Italy), an institute for which we jointly produced a report covering an overview of the different theories developed around the subject, as well as the main tools proposing solutions to this problem. The aim of this collaboration is at understanding the basic principles and the computer programs apt to coordinate a public debate with an overall aim at giving the bases for designing such programs. Computer programs for the coordination of public debate exist since the beginning of the eighties but recently they have acquired new relevance for the use made of them by public administrations, associations and political parties. The meet of both citizen's needs and public administrations for transparency can today be technically realized with such programs through the present communication means in a more efficient way with respect to the first experiments dating now about forty years. This report aims at covering historical, technical and some theoretical aspects of the use of computers for the coordination of public debate.

7.4.2.3. Orthomodular partial orders.

The collaboration with Carlo Ferigato, is in line with the latter's thesis subject [11]. The set of regions of a condition/event transition system represents all the possible local states of a net system the behaviour of which is specified by the transition system. This set can be endowed with a structure, so as to form an orthomodular partial order. Given such a structure, one can then define another condition/event transition system. We study cases in which this second transition system has the same collection of regions as the first one. When it is so, the structure of regions is called stable. We proposed, to this aim, a composition operation, and a refinement operation for stable orthomodular partial orders, the results of which are stable.

7.5. New results on Transversal Concern: Missing Models

Participants : Hugo Bazille, Sihem Cherrared, Éric Fabre, Blaise Genest, Thierry Jéron, The Anh Pham

7.5.1. Unfolding-based dynamic partial-order reduction of asynchronous distributed programs

Unfolding-based Dynamic Partial Order Reduction (UDPOR) is a recent technique mixing Dynamic Partial Order Reduction (DPOR) with concepts of concurrency such as unfoldings to efficiently mitigate state space explosion in model-checking of concurrent programs. It is optimal in the sense that each Mazurkiewicz trace, *i.e.* a class of interleavings equivalent by commuting independent actions, is explored exactly once. In this work [25] we show that UDPOR can be extended to verify asynchronous distributed applications, where processes both communicate by messages and synchronize on shared resources. To do so, a general model

of asynchronous distributed programs is formalized in TLA+. This allows to define an independence relation, a main ingredient of the unfolding semantics used by UDPOR during the UDPOR exploration. Then, the adaptation of UDPOR, involving the construction of an unfolding during the execution of the applicaton (*i.e.* with no model of the application but the code iteself), is made efficient by a precise analysis of dependencies. A prototype implementation gives promising experimental results.

7.5.2. Learning models for telecommunication management.

Model based methods have been recognised as the most appropriate approach to fault diagnosis in telecommunication networks, as they not only help in detecting and classifying failures, but is also provides useful explanations about the propagation of faults in such large distributed and concurrent systems. However, the bottleneck of these methods is of course the derivation and validation of a relevant model [8]. We have explored two techniques in this direction, based on fault/stress injection.

A first approach (collaboration Orange Labs) [33] consists in assembling generic components that would match the current (changing) topology of a software defined network. The model can then be validated by fault injection on a platform running the true VNF (virtual network functions) chains that are used in production. The second approach (collaboration Nokia Bell Labs) aims at detecting soft performance degradations, that would impact the quality of service, but not produce faults and alarms. Again, this can be achieved by stress injection at the level of VMs (virtual machines) in production software, and by collecting signature patterns under the form of statistical changes in the performance metrics collected on such systems.

7.5.3. Verification of deep neural networks.

Deep neural networks are as effective in their respective tasks as hardly understandable by a human. To use them in critical applications, not only they should be understood, they must be certified. We surveyed in [14] a large number of recent attempts to formally certify deep neural networks obtained by deep machine learning techniques. Most of the work currently focus on forward-propagating networks, and the problem of certifying their robustness.

TAMIS Project-Team

6. New Results

6.1. Results for Axis 1: Vulnerability analysis

6.1.1. New Advances on Side-channel Distinguishers

Participants: Christophe Genevey Metat, Annelie Heuser.

A Systematic Evaluation of Profiling Through Focused Feature Selection.
Profiled side-channel attacks consist of several steps one needs to take. An important, but sometimes ignored, step is a selection of the points of interest (features) within side-channel measurement traces. A large majority of the related works start the analyses with an assumption that the features are preselected. Contrary to this assumption, here, we concentrate on the feature selection step. We investigate how advanced feature selection techniques stemming from the machine learning domain can be used to improve the attack efficiency. To this end, we provide a systematic evaluation of the methods of interest. The experiments are performed on several real-world data sets containing software and hardware implementations of AES, including the random delay countermeasure. Our results show that wrapper and hybrid feature selection methods perform extremely well over a wide range of test scenarios and a number of features selected. We emphasize L1 regularization (wrapper approach) and linear support vector machine (SVM) with recursive feature elimination used after chi-square filter (Hybrid approach) that performs well in both accuracy and guessing entropy. Finally, we show that the use of appropriate feature selection techniques is more important for an attack on the high-noise data sets, including those with countermeasures, than on the low-noise ones.

[3] Make Some Noise. Unleashing the Power of Convolutional Neural Networks for Profiled Side-channel Analysis. *Profiled side-channel analysis based on deep learning, and more precisely Convolutional Neural Networks, is a paradigm showing significant potential. The results, although scarce for now, suggest that such techniques are even able to break cryptographic implementations protected with countermeasures. In this paper, we start by proposing a new Convolutional Neural Network instance able to reach high performance for a number of considered datasets. We compare our neural network with the one designed for a particular dataset with masking countermeasure and we show that both are good designs but also that neither can be considered as a superior to the other one. Next, we address how the addition of artificial noise to the input signal can be actually beneficial to the performance of the neural network. Such noise addition is equivalent to the regularization term in the objective function. By using this technique, we are able to reduce the number of measurements needed to reveal the secret key by orders of magnitude for both neural networks. Our new convolutional neural network instance with added noise is able to break the implementation protected with the random delay countermeasure by using only 3 traces in the attack phase. To further strengthen our experimental results, we investigate the performance with a varying number of training samples, noise levels, and epochs. Our findings show that adding noise is beneficial throughout all training set sizes and epochs.*

The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-channel Evaluations.

We concentrate on machine learning techniques used for profiled sidechannel analysis in the presence of imbalanced data. Such scenarios are realistic and often occurring, for instance in the Hamming weight or Hamming distance leakage models. In order to deal with the imbalanced data, we use various balancing techniques and we show that most of them help in mounting successful attacks when the data is highly imbalanced. Especially, the results with the SMOTE technique are encouraging, since we observe some scenarios where it reduces the number of necessary measurements more than 8 times. Next, we provide extensive results on comparison of machine learning and side-channel metrics, where we show that machine learning metrics (and especially accuracy as the most often used one) can be extremely deceptive. This finding opens a need to revisit the previous works and their results in order to properly assess the performance of machine learning in side-channel analysis.

[5] CC Meets FIPS: A Hybrid Test Methodology for First Order Side Channel Analysis.
Common Criteria (CC) and FIPS 140-3 are two popular side channel testing methodologies. Test Vector Leakage Assessment Methodology (TVLA), a potential candidate for FIPS, can detect the presence of side-channel information in leakage measurements. However, TVLA results cannot be used to quantify side-channel vulnerability and it is an open problem to derive its relationship with side channel attack success rate (SR), i.e., a common metric for CC. In this paper, we extend the TVLA testing beyond its current scope. Precisely, we derive a concrete relationship between TVLA and signal to noise ratio (SNR). The linking of the two metrics allows direct computation of success rate (SR) from TVLA for given choice of intermediate variable and leakage model and thus unify these popular side channel detection and evaluation metrics. An end-to-end methodology is proposed, which can be easily automated, to derive attack SR starting from TVLA testing. The methodology works under both univariate and multivariate setting and is capable of quantifying any first order leakage. Detailed experiments have been provided using both simulated traces and real traces on SAKURA-GW platform. Additionally, the proposed methodology is benchmarked against previously published attacks on DPA contest v4.0 traces, followed by extension to jitter based countermeasure. The result shows that the proposed methodology provides a quick estimate of SR without performing actual attacks, thus bridging the gap between CC and FIPS.

[13] Combining sources of side-channel information.
A few papers relate that multi-channel consideration can be beneficial for side-channel analysis. However, all were conducted using classical attack techniques. In this work, we propose to explore a multi-channel approach thanks to machine/deep learning. We investigate two kinds of multi-channel combinations. Unlike previous works, we investigate the combination of EM emissions from different locations capturing data-dependent leakage information on the device. Additionally, we consider the combination of the classical leaking signals and a measure of mostly the ambient noise. The knowledge of the ambient noise (due to WiFi, GSM, ...) may help to remove it from a noisy trace. To investigate these multi-channel approaches, we describe one option of how to extend a CNN architecture which takes as input multiple channels. Our results show that multi-channel networks are suitable for side-channel analysis. However, if one channel alone already contains enough exploitable information to reach high effectiveness, naturally, the multi-channel approach cannot improve the performance further.

6.1.2. Side-channel analysis on post-quantum cryptography

Participants: Tania Richmond, Yulliwas Ameur, Agathe Cheriere, Annelie Heuser.

In recent years, there has been a substantial amount of research on quantum computers ? machines that exploit quantum mechanical phenomena to solve mathematical problems that are difficult or intractable for conventional computers. If large-scale quantum computers are ever built, they will be able to break many of the public-key cryptosystems currently in use. This would seriously compromise the confidentiality and integrity of digital communications on the Internet and elsewhere. The goal of post-quantum cryptography (also called quantum-resistant cryptography) is to develop cryptographic systems that are secure against both quantum and classical computers, and can interoperate with existing communications protocols and networks. At present,

there are several post-quantum cryptosystems that have been proposed: lattice-based, code-based, multivariate cryptosystems, hash-based signatures, and others. However, for most of these proposals, further research is needed in order to gain more confidence in their security and to improve their performance. Our interest lies in particular on the side-channel analysis and resistance of these post-quantum schemes, in particular code-based cryptosystems.

During this year, we have set up a first side-channel experiment platform suited for embedded devices running code-based cryptosystems. Using this platform we exploited vulnerabilities of the syndrome computation present in some code-based algorithms.

6.1.3. Verification of IKEv2 protocol

Participants: Tristan Ninet, Olivier Zendra.

The IKEv2 (Internet Key Exchange version 2) protocol is the authenticated key-exchange protocol used to set up secure communications in an IPsec (Internet Protocol security) architecture. IKEv2 guarantees security properties like mutual-authentication and secrecy of exchanged key. To obtain an IKEv2 implementation as secure as possible, we use model checking to verify the properties on the protocol specification, and software formal verification tools to detect implementation flaws like buffer overflows or memory leaks.

In previous analyses, IKEv2 has been shown to possess two authentication vulnerabilities that were considered not exploitable. We analyze the protocol specification using the Spin model checker, and prove that in fact the first vulnerability does not exist. In addition, we show that the second vulnerability is exploitable by designing and implementing a novel slow Denial-of-Service attack, which we name the Deviation Attack.

We propose an expression of the time at which Denial-of-Service happens, and validate it through experiment on the strongSwan implementation of IKEv2. As a counter-measure, we propose a modification of IKEv2, and use model checking to prove that the modified version is secure.

For ethical reasons we informed our country's national security agency (ANSSI) about the existence of the Deviation Attack. The security agency gave us some technical feedback as well as its approval for publishing the attack.

We then tackle formal verification directly applied to an IKEv2 source code. We already tried to analyze strongSwan using the Angr tool. However we found that the Angr was not mature yet for a program like strongSwan. We thus try other software formal verification tools and apply them to smaller and simpler source code than strongSwan: we analyze OpenSSL asn1parse using the CBMC tool and light-weight IP using the Infer tool. We find that CBMC does not scale to a large source code and that Infer does not verify the properties we want.

We explored more in-depth a formal technique and work towards the goal of verifying generic properties (absence of implementation flaws) on softwares like strongSwan.

Publications:

- [10] Model Checking the IKEv2 Protocol Using Spin
- [11] The Deviation Attack: A Novel Denial-of-Service Attack Against IKEv2

6.1.4. Software obfuscation

Participants: Alexandre Gonzalvez, Olivier Decourbe.

The limits of software obfuscation are not clear in practice. A protection based on opaque predicates can not be compatible with the control flow integrity property at low-level, due to the presence of indirect jumps in the instruction set architecture semantics. We propose a restricted instruction set architecture to overcome this limit. We argue for the adoption of restricted instruction set architecture for security-related computation. Publication:

- [9] A case against indirect jumps for secure programs

6.2. Results for Axis 2: Malware analysis

The detection of malicious programs is a fundamental step to be able to guarantee system security. Programs that exhibit malicious behavior, or *malware*, are commonly used in all sort of cyberattacks. They can be used to gain remote access on a system, spy on its users, exfiltrate and modify data, execute denial of services attacks, etc.

Significant efforts are being undertaken by software and data companies and researchers to protect systems, locate infections, and reverse damage inflicted by malware. Our contribution to malware analysis include the following fields:

6.2.1. Malware Classification and clustering

Participants: Cassius Puodzius, Stefano Sebastio, Olivier Decourbe, Annelie Heuser, Olivier Zendra.

Once malicious behavior has been located, it is essential to be able to classify the malware in its specific family to know how to disinfect the system and reverse the damage inflicted on it.

While it is rare to find an actually previously unknown malware, morphic techniques are employed by malware creators to ensure that different generations of the same malware behave differently enough than it is hard to recognize them as belonging to the same family. In particular, techniques based on the syntax of the program fails against morphic malware, since syntax can be easily changed.

To this end, semantic signatures are used to classify malware in the appropriate family. Semantic signatures capture the malware's behavior, and are thus resistant to morphic and differentiation techniques that modify the malware's syntactic signatures. We are investigating semantic signatures based on the program's System Call Dependency Graph (SCDG), which have been proven to be effective and compact enough to be used in practice. SCDGs are often extracted using a technique based on pushdown automata that is ineffective against obfuscated code; instead, we are applying concolic analysis via the `angr` engine to improve speed and coverage of the extraction.

Once a semantic signature has been extracted, it has to be compared against large database of known signatures representing the various malware families to classify it. The most efficient way to obtain this is to use a supervised machine learning classifier. In this approach, the classifier is trained with a large sample of signatures malware annotated with the appropriate information about the malware families, so that it can learn to quickly and automatically classify signatures in the appropriate family. Our work on machine learning classification focuses on using SCDGs as signatures. Since SCDGs are graphs, we are investigating and adapting algorithms for the machine learning classification of graphs, usually based on measures of shared subgraphs between different graphs. One of our analysis techniques relies on common subgraph extraction, with the idea that a malicious behavior characteristic of a malware family will yield a set of common subgraphs. Another approach relies on the Weisfeiler-Lehman graph kernel which uses the presence of nodes and their neighborhoods pattern to evaluate similarity between graphs. The presence or not of a given pattern becomes a feature in a subsequent machine learning analysis through random forest or SVM.

Moreover, we explored the impact on the malware classification of several heuristics adoptable in the SCDGs building process and graph exploration. In particular, our purpose was to:

- identify quality characteristics and evaluation metrics of binary signatures based on SCDGs (and consequently the key properties of the execution traces), that characterize signatures able to provide high-precision malware classification
- optimize the performance of the SMT solver by designing a meta-heuristic able to select the best heuristic to tackle a specific sub-class of problem, study the impact of the configuration of the SMT solver and symbolic execution framework, and understand their interdependencies with the aim of efficiently extracting SCDGs in accordance with the identified quality metrics.

By adopting a Design of Experiments approach constituted by a full factorial experiment design and an Analysis of Variance (ANOVA) we have been able to pinpoint that, considering the graph metrics and their impact on the F-score, the litmus test for the quality of an SCDG-based classifier is represented by the presence of connected components. This could be explained considering how the graph mining algorithm (gSpan) works and the adopted similarity metric based on the number of common edges between the extracted signatures and the SCDG of the sample to classify. The results of the factorial experiments show that in our context tuning the symbolic execution is a very complex problem and that the sparsity of effect principle (stating that the system is dominated by the effect of the main factors and low-order-factor interactions) does not hold. The evaluation proved that the SMT solver is the most influential positive factor also showing an ability in reducing the impact of heuristics that may need to be enabled due to resource constraints (e.g., the max number of active paths). Results suggest that the most important factors are the disjoint union (as trace combination heuristic), and the our SMT optimization (through meta-heuristics) whereas other heuristics (such as min trace size and step timeout) have less impact on the quality of the constructed SCDGs.

During this year we build a end-to-end functional toolchain for supervised learning.

Furthermore, we have extended our approach to malware classification using unsupervised clustering. Preliminary results show that we are able to classify malware according to their behavioral properties without the need of any predefined labels.

6.2.2. Packers analysis

Participants: Lamine Nourredine, Cassius Puodzius, Stefano Sebastio, Annelie Heuser, Olivier Zendra.

Packing is a widespread tool to prevent static malware detection and analysis. Detecting and classifying the packer used by a given malware sample is fundamental to being able to unpack and study the malware, whether manually or automatically. Existing works on packing detection and classification has focused on effectiveness, but does not consider the efficiency required to be part of a practical malware-analysis workflow. This work studies how to train packing detection and classification algorithms based on machine learning to be both highly effective and efficient. Initially, we create ground truths by labeling more than 280,000 samples with three different techniques. Then we perform feature selection considering the contribution and computation cost of features. Then we iterate over more than 1,500 combinations of features, scenarios, and algorithms to determine which algorithms are the most effective and efficient, finding that a reduction of 1-2% effectiveness can increase efficiency by 17-44 times. Then, we test how the best algorithms perform against malware collected after the training data to assess them against new packing techniques and versions, finding a large impact of the ground truth used on algorithm robustness. Finally, we perform an economic analysis and find simple algorithms with small feature sets to be more economical than complex algorithms with large feature sets based on uptime/training time ratio.

A limit of supervised learning is to not be able to recognize classes that were not present in the ground truth. In the work's case above, this means that packer families for which a classifier has not been trained will not be recognized. In this work, we use unsupervised learning techniques, more particularly clustering, in order to provide information about packed malware with previously unknown packing techniques. Here, we build our own dataset of packed binaries, since in the previous work, it has been shown that the construction of the ground truth was fundamental in determining the effectiveness of the packing classification process. Choosing the right clustering algorithm with the right distance metric, dealing with different scales of features units, while being effective, efficient and robust are also majors parts of the current work.

During this year we have developed a toolchain of effective clustering of packers, in particular taking into account the possibility of evolution in packers. For this we derived and implemented new feature extraction strategies combined with incremental clustering algorithms.

6.3. (Coordination of the) H2020 TeamPlay Project, and Expression of Security Properties

Participants: Olivier Zendra, Yoann Marquer, Céline Minh, Nicolas Kiss, Annelie Heuser, Tania Richmond.

6.3.1. Overview & results

This work is done in the context of the TeamPlay EU project.

As mobile applications, the Internet of Things, and cyber-physical systems become more prevalent, so there is an increasing focus on energy efficiency of multicore computing applications. At the same time, traditional performance issues remain equally important. Increasingly, software designs need to find the best performance within some energy budget, often while also respecting real-time or other constraints, which may include security, data locality or system criticality, and while simultaneously optimising the usage of the available hardware resources.

While parallel multicore/manycore hardware can, in principle, ameliorate energy problems, and heterogeneous systems can help to find a good balance between execution time and energy usage, at present there are no effective analyses beyond user-guided simulations that can reliably predict energy usage for parallel systems, whether alone or in combination with timing information and security properties. In order to create energy-, time- and security- (ETS) efficient parallel software, programmers need to be actively engaged in decisions about energy usage, execution time and security properties rather than passively informed about their effects. This extends to design-time as well as to implementation-time and run-time.

In order to address this fundamental challenge, TeamPlay takes a radically new approach: by exploiting new and emerging ideas that allow non-functional properties to be deeply embedded within their programs, programmers can be empowered to directly treat energy ETS properties as first-class citizens in their parallel software. The concrete objectives of the TeamPlay project are:

1. To develop new mechanisms, along with their theoretical and practical underpinnings, that support direct language-level reasoning about energy usage, timing behaviour, security, etc.
2. To develop system-level coordination mechanisms that facilitate optimised resource usage for multicore hardware, combining system-level resource utilisation control during software development with efficient spatial and temporal scheduling at run-time.
3. To determine the fundamental inter-relationships between time, energy, security, etc. optimisations, to establish which optimisation approaches are most effective for which criteria, and to consequently develop multiobjective optimising compilers that can balance energy consumption against timing and other constraints.
4. To develop energy models for heterogeneous multicore architectures that are sufficiently accurate to enable high-level reasoning and optimisation during system development and at run-time.
5. To develop static and dynamic analyses that are capable of determining accurate time, energy usage and security information for code fragments in a way that can inform high-level programs, so achieving energy, time and security transparency at the source code level.
6. To integrate these models, analyses and tools into an analysis-based toolbox that is capable of reflecting accurate static and dynamic information on execution time and energy consumption to the programmer and that is capable of optimising time, energy, security and other required metrics at the whole system level.
7. To identify industrially-relevant metrics and requirements and to evaluate the effectiveness and potential of our research using these metrics and requirements.
8. To promote the adoption of advanced energy-, time- and security-aware software engineering techniques and tools among the relevant stake-holders.

Inria will exploit the results of the TeamPlay project in two main domains. First, they will strengthen and extend the research Inria has been carrying on low power and energy for embedded systems, especially for memory and wireless sensors networks. Second, they will complement in a very fitting way the research carried at Inria about security at a higher level (model checking, information theory).

The capability to express the energy and security properties at the developper level will be integrate in Inria own prototype tools, hence widening their applicability and the ease of experimentation. The use of energy properties wrt. evening of energy consumption to prevent information leakage, thus making side-channels attacks more difficult, is also a very promising path.

In addition, the methodological results pertaining to the development of embedded systems with a focus on low power and energy should also contribute to research lead at Inria in the domain of software engineering and advanced software engineering tools. Furthermore, security research lead at Inria will benefit from the security work undertaken by Inria and SIC in TeamPlay.

Overall, the project, with a strong industrial presence, will allow Inria to focus on matching concrete industrial requirements aiming at actual products, hence in providing more robust and validated results. In addition, the extra experience of working with industrial partners including SMEs will surely impact positively on Inria research methodology, making Inria research more attractive and influential, especially wrt. industry.

Finally, the results, both in terms of methodology and techniques, will also be integrated in the teaching Inria contributes to at Master level, in the areas of Embedded Systems and of Security.

The TeamPlay consortium agreement has been created by Inria, discussed with the various partners, and has been signed by all partners on 28 Feb. 2018. Inria has also distributed the partners initial share of the grant at the beginning of the project.

As WP7 (project management) leader and project coordinator, Inria was in charge of arranging general project meetings, including monthly meetings (tele-conferences), bi-annual physical meetings, boards meetings. During the first period, three exceptional physical meetings have been conducted, in addition to monthly project meetings: the kick-off meeting in Rennes from the 30th to the 31st of January 2018, the physical progress meeting has been conducted in Odense from the 26th to the 27th of June 2018, and the review in Brussels prepared the 19th of September 2018 and set the 17th of October 2018.

We have selected and set up utility tools for TeamPlay: shared notepads, mailing lists, shared calendars and collaborative repositories. We have ensured the timely production of the due deliverables. We set up the Project Advisory Board (PAB) with the aim of gathering external experts from both academia and industry, covering a wide range of domains addressed by TeamPlay. Finally, we ensured good working relationships (which can implicate conflict resolution when needed), monitored the overall progress of the project, and reported to the European Commission on technical matters and deliverables.

We also organized a tooling meeting in Hamburg in October the 30th, to discuss the relation between the tools from different partners, e.g. Idris from the University of St Andrews, the WCC compiler developed in the Hamburg University of Technology, or the coordination tool developed in the University of Amsterdam.

Measuring security, unlike measuring other more common non-functional properties like time or energy, is still very much in its infancy. For example, time is often measured in seconds (or divisions thereof), but security has no widely agreed, well-defined measurement. It is thus one goal of this project, especially for SIC and Inria, to design (necessarily novel) security measurements, and have them implemented as much as possible throughout the set of development tools.

Measuring security by only one value however seems impossible or may be meaningless. More precisely, if security could be defined overall by only one measurement, the latter would be a compound (i.e. an aggregation) of several more specialized measurement. Indeed, security encompasses many aspects of interest:

1. By allowing communications between different systems, security properties should be guaranteed in order to prevent low-level users from determining anything about high-level users activity, or in the case of public communication channels in a hostile environment, to evaluate vulnerability to intruders performing attacks on communications.
    1. *Confidentiality* (sometimes called *secrecy*) properties like non-interference (and many) variants can be described by using an information-flow policy (e.g. high- and low-level users) and studying traces of user inputs.
    2. *Vulnerability* captures how a system is sensible to attacks on communications (e.g. stealing or faking information on a public channel).
2. A *side-channel* is a way of transmitting informations (purposely or not) to another system out of the standard (intended) communication channels. *Side-channel attacks* rely on the relationship between information leaked through a side-channel and the secret data to obtain confidential (non-public) information.

1. *Entropy* captures the uncertainty of the attacker about the secret key. The attacker must be able to extract information about the secret key through side-channel measurements, which is captured by the *attacker's remaining uncertainty* value, which can be computed by using heuristic techniques. The attacker must also be able to effectively recover the key from the extracted information, which is expressed by the *min-entropy leakage*, and refined by the *g-leakage* of a gain function.
2. The power consumption of a cryptographic device can be analyzed to extract the secret key. This is done by using several techniques: visual examination of graphs of the current (*Simple Power Analysis*), by exploiting biases in varying power consumption (*Differential Power Analysis*), or by using the correlation coefficient between the power samples and hypotheses (*Correlation Power Analysis*).
3. Usual security properties guarantee only the input-output behavior of a program, and not its execution time. Closing *leakage through timing* can be done by disallowing while-loops and if-commands to depend on high security data, or by padding the branches so that the external observer cannot determine which branch was taken.
4. Finally, the correlation between the patterns of the victim's execution and the attacker's observations is formalized as a metric called the *Side-channel Vulnerability Factor*, which is refined by the *Cache Side-channel Vulnerability* for cache attacks.

3. A cryptographic scheme should be secure even if the attacker knows all details about the system, with the exception of the secret keys. In particular, the system should be secure when the attacker knows the encryption and decryption algorithms.
    1. In modern cryptography, the security level (or security strength) is given by the *work factor*, which is related to its key-length and the number of operations necessary to break a cryptographic scheme (try all possible combinations of the key). An algorithm is said to have a "security level of n bits" if the best known attack requires 2^n steps. This is a quite natural definition because symmetric algorithms with a security level of n have a key of length n bits.
    2. The relationship between cryptographic strength and security is not as straightforward in the asymmetric case. Moreover, for symmetric algorithms, a key-length of 128 bits provides an estimated long term security (i.e. several decades in the absence of quantum computer) regarding brute-force attacks. To reach an estimated long term security even with quantum computers, a key-length of 256 bits is mandatory.

Inria is implementing side-channel countermeasures (hiding) into the WCET-aware C Compiler (WCC) developed by the Hamburg University of Technology (TUHH). A research visit to TUHH was arranged with the aim at learning how to work on WCC (TUHH and WCC infrastructure, WCC developers best practices, etc.). Inria will use compiler-based techniques to prevent timing leakages and power leakages.

For instance, in a conditional branching `if` b `then` $P_1(x)$ `else` $P_2(x)$, measuring the execution time or the power profile may allow to know whether the branch P_1 or P_2 have been chosen to manipulate the value x, thus to obtain the secret value b. To prevent timing leakage, P_1 and/or P_2 can be padded (i.e. dummy instructions are added) in order to obtain the worst-case execution time in both branches.

But this does not prevent information leakage from power profile. A stronger technique, from a security point of view, could be to add a dummy variable y and duplicate the code such that $y = x$; `if` b `then` $P_1(x); P_2(y)$ `else` $P_1(Y); P_2(x)$ always performs the operations of P_1 then the operations of P_2. But the execution time is now the sum and not the worst-case of both branches, thus trading execution time to increase security.

Finally, the initialization $y = x$ can be detected, and the previous solution is still vulnerable to fault injections. Some algorithms like the Montgomery Ladder are more protected against these attacks because both variables x and y are entangled during the execution. We hope to generalize this property to a wider set of algorithms, or to automatically detect the properties required from the original code in order to transform it into a “ladderized" version with higher security level.

6.3.2. Publication

- Type-Driven Verification of Non-functional Properties [8].

TEA Project-Team

7. New Results

7.1. ADFG: Affine data-flow graphs scheduler synthesis

Participants: Loïc Besnard, Thierry Gautier, Jean-Pierre Talpin, Shuvra Bhattacharyya, Alexandre Honorat, Hai Nam Tran.

ADFG (Affine DataFlow Graph) synthesizes scheduling parameters for real-time systems modeled as synchronous data flow (SDF), cyclo-static dataflow (CSDF), and ultimately cyclo-static dataflow (UCSDF) graphs. It aims at mitigating the trade-off between throughput maximization and total buffer size minimization. The synthesizer inputs are a graph which describes tasks by their Worst Case Execution Time (WCET), and directed buffers connecting tasks by their data production and consumption rates; the number of processors in the target system and the real-time scheduling synthesis algorithm to be used. The outputs are synthesized scheduling parameters such as tasks periods, offsets, processor bindings, priorities, buffer initial markings and buffer sizes. ADFG was originally implemented by Adnan Bouakaz [0]. It is now being collaboratively developed with team Tea, Hai Nam Tran (UBO) Alexandre Honorat (INSA) and Shuvra Bhattacharyya (UMD/INSA/Inria).

ADFG is extended to support automated code generation of the computed buffer sizes and scheduling parameters for dataflow applications that are implemented in the Lightweight Dataflow Environment (LIDE) [0]. LIDE is a flexible, lightweight design environment that allows designers to experiment with dataflow-based implementations directly. LIDE actors and buffers (FIFOs) can be initialized with parameters, including buffer sizes. The usage of LIDE allows a systematic way to instantiate dataflow graphs with the buffer size parameters computed by ADFG.

Actor models and scheduling algorithms in ADFG have been extended to investigate the contention-aware scheduling problem on multi/many-core architectures. The problem we tackled is that the scheduler synthesis for these platforms must account for the non-negligible delay due to shared memory accesses. We exploited the deterministic communications exposed in SDF graphs to account for the contention and further optimize the synthesized schedule. Two solutions are proposed and implemented in ADFG: contention-aware and contention-free scheduling synthesis. In other words, we either take into account the contention and synthesize a contention-aware schedule or find a one that results in no contention.

ADFG is extended to apply a transformation known as partial expansion graphs (PEG). This transformation can be applied as a pre-processing stage to improve the exploitation of data parallelism in SDF graphs on parallel platforms. In contrast to the classical approaches of transforming SDF graphs into equivalent homogeneous forms, which could lead to an exponential increase in the number of actors and excessive communication overhead, PEG-based approaches allow the designer to control the degree to which each actor is expanded. A PEG algorithm that employs cyclo-static data flow techniques is developed in ADFG. Compared to existing PEG-based approach, our solution requires neither buffer managers nor split-join actors to coordinate data production and consumption rates. This allows us to reduce the number of added actors and communication overhead in the expanded graphs.

7.2. Parallel Composition and Modular Verification of Computer Controlled Systems in Differential Dynamic Logic

Participants: Jean-Pierre Talpin, Benoit Boyer, David Mentre, Simon Lunel, Stefan Mitsch.

[0]Real-Time Scheduling of Dataflow Graphs. A. Bouakaz. Ph.D. Thesis, University of Rennes 1, 2013.

[0]S. Lin, Y. Liu, K. Lee, L. Li, W. Plishker, and S. S. Bhattacharyya. 2017. The DSPCAD framework for modeling and synthesis of signal processing systems. Handbook of Hardware/Software Codesign (2017), 1185–1219.

The primary goal of our project, in collaboration with Mitsubishi Electronics Research Centre Europe (MERCE), is to ensure correctness-by-design in realistic cyber-physical systems, i.e., systems that mix software and hardware in a physical environment, e.g., Mitsubishi factory automation lines or water-plant factory. To achieve that, we develop a verification methodology based on the decomposition of systems into components enhanced with compositional contract reasoning.

The work of A. Platzer on Differential Dynamic Logic ($d\mathcal{L}$) held our attention [0]. This formalism is built upon the Dynamic Logic of V. Pratt and augmented with the possibility of expressing Ordinary Differential Equations (ODEs). Combined with the ability of Dynamic Logic to specify and verify hybrid programs, $d\mathcal{L}$ is particularly adapted to model cyber-physical systems. The proof system associated with the logic is implemented into the theorem prover KeYmaera X. Aimed toward automation, it is a promising tool to spread formal methods in industry.

Computer-Controlled Systems (CCS) are a subclass of hybrid systems where the periodic relation of control components to time is of paramount importance. Since they additionally are at the heart of many safety-critical devices, it is of primary importance to correctly model such systems and to ensure they function correctly according to safety requirements. Differential dynamic logic $d\mathcal{L}$ is a powerful logic to model hybrid systems and to prove their correctness. We contributed a compositional modeling and reasoning framework to $d\mathcal{L}$ that separates models into components with timing guarantees, such as reactivity of controllers and controllability of continuous dynamics. Components operate in parallel, with coarse-grained interleaving, periodic execution and communication. We present techniques to automate system safety proofs from isolated, modular, and possibly mechanized proofs of component properties parameterized with timing characteristics.

7.3. Multithreaded code generation for process networks

Participants: Loïc Besnard, Thierry Gautier.

As part of an in-depth comparison of process models, we have recently revisited the relation between the model of asynchronous dataflow represented by Kahn Process Networks (KPNs) and that of synchronous dataflow represented by the polychronous model of computation. In particular, we have precisely described in which conditions polychronous programs can be seen as KPNs. In this context, we have considered different cases of process networks, including so-called "polyendochronous processes". Under some conditions expressed by clock equation systems, (networks of) processes exhibiting polyhierarchies of clocks are polyendochronous and, as compositions of endochronous processes, may be seen as KPNs.

Based on this characterization, we have developed in the open-source Polychrony toolset a new strategy of code generation for such (polyendochronous) process networks. Typically, after the clock calculus, a program P is organized as a composition of processes, $P = (|\ P1 \mid P2 \mid ... \mid Pn\ |)$, each one structured around a clock tree. When P is characterized as polyendochronous, it contains generally clock constraints such as $Clk1 = Clk2$, with $Clk1$ being a clock in the subtree corresponding to $P1$ and $Clk2$ a clock in the subtree corresponding to $P2$.

Such a constraint induces a synchronization between two parts ($P1$, $P2$) of the program when $Clk1$ or $Clk2$ occurs. The principle of the code generation for polyendochronous processes is based on the existing distributed code generation, but with the additional resynchronization of parts of the application induced by the constraints on clocks ($Clk1$, $Clk2$) not placed in the same clock trees. For distributed code generation, it is considered that each (clock) hierarchy will run on a specific processor. In this case, the purpose is mainly to partition the application, and the processors will be virtual ones.

The code generation of each partition consists in the definition of several tasks: one task per cluster (a cluster being a subpart that may be executed as soon as its inputs are available, without any communication with the external world); one task per input/output of the partition; one task for the cluster of state variables; one task that manages the steps. Synchronization between these tasks is obtained by semaphores (one semaphore per task). This code generation technique for the class of networks called "polyendochronous processes" has been

[0] *Differential Dynamic Logic for Hybrid Systems*, André Platzer, http://symbolaris.com/logic/dL.html

added in the Polychrony toolset (http://polychrony.inria.fr) and a paper describing the comparison of process models is currently in submission.

7.4. Type theory for modular static analysis of system programs

Participants: Lucas Franceschino, Jean-Pierre Talpin, David Pichardie.

This Ph.D. project is about formal verification, with system programming applications in mind. Formal methods are essential for safety-critical software (i.e. transport and aeronautic industry). In the same time, more and more programming languages with a strong type system arise (such as Haskell, Rust, ML, Coq, F*, Idris...).

Formal methods come in different flavors: type theory, abstract interpretation, refinement types. Each of these "flavors" are both theoretical fields and are also being implemented concretely: *Astrée* ou *Verasco* for abstract interpretation, *Coq*, *Agda*, *F** or *Idris* dependent types, and *Liquid Haskell* for refinement types.

Our approach consists in positioning ourselves between type theory and abstract interpretation, and to leverage the power of both. The main intuition behind this idea is that abstract interpretation, suffering from expressiveness, would bring *invariant inference* power, while strong type systems, requiring manual annotations and proofs, would bring *expressivity*.

We formalized how one can enrich a weakest precondition calculus (WP) with an abstract interpreter. This work takes the shape of a WP calculus transformer: given a WP calculus, we generically construct a brand new WP calculus that produces easier (but sound, still) weakest preconditions, thanks to abstract interpretation.

Concretely, our work is being implemented as an F* effect transformer that leverage Verasco capabilities, for a low-level subset of F*, namely Low*.

7.5. Verified information flow of embedded programs

Participants: Jean-Joseph Marty, Lucas Franceschino, Niki Vazou, Jean-Pierre Talpin.

This PhD project is about applying refinement types theory to verified programming of applications and modules of library operating systems, such as unikernels, for embedded devices of the Internet of Things (IoT): TinyOS, Riot, etc. Our topic has focused on developing a model of information flow control using labeled input-outputs (LIO) implemented using F☆: project Lio☆.

As part of the development of Lio☆, we implemented a library that, thanks to static verification, ensures the containment of information in relation to a parameterized policy for information flow control. In collaboration with Niki Vazou (IMDEA) and Lucas Franceschino we have formalized and developed an automatic method to prove non-interference in Meta☆. Using the Kremlin code generator, programs using Lio☆ can be compiled into C code and run natively on embedded low-resource-constrained devices, without the need for additional runtime system.

In parallel we continued our collaboration with the ProgSys team on a second, now discontinued, project: Gluco☆. The goal of this project was to evaluate the capabilities to use the F* programming language to program an entire system by taking into account its software, hardware and physical constraints using type refinements [0].

[0] Towards verified programming of embedded devices. J.-P. Talpin, J.-J. Marty, S. Narayan, D. Stefan, R. Gupta. Design, Automation and Test in Europe (DATE'19). IEEE, 2019.

I4S Project-Team

6. New Results

6.1. System identification

6.1.1. On Local LTI Model Coherence for LPV Interpolation

Participant: Qinghua Zhang.

In the local approach to linear parameter varying (LPV) system identification, it is widely acknowledged that locally estimated linear state-space models should be made coherent before being interpolated, but the accurate meaning of the term " coherent " or " coherence " is rarely defined. The purpose of this study is to analyze the relevance of two existing definitions and to point out the consequence of this analysis on the practice of LPV system identification. This work has been carried out in collaboration with Lennart Ljung of Linköping University and Rik Pintelon of Vriije Universiteit Brussel, and the results have been published in [22].

6.1.2. Stability Analysis of the Kalman Predictor

Participant: Qinghua Zhang.

The stability of the Kalman filter, though less often mentioned than the optimality in the recent literature, is a crucial property for real time applications. The purpose of this paper is to complete the classical stability analysis of the Kalman filter for general time varying systems. A proof of the stability of the one step ahead predictor, which is embedded in the Kalman filter, is presented in this paper, whereas the classical results were focused on the stability of the filter. The predictor stability is particularly important for linear parameter varying (LPV) system identification by means of prediction error minimization. This work has been carried out in collaboration with Liangquan Zhang of Beijing University of Posts and Telecommunications, and the results have been published in [23].

6.1.3. Regularized Adaptive Observer to Address Deficient Excitation

Participant: Qinghua Zhang.

Adaptive observers are recursive algorithms for joint estimation of both state variables and unknown parameters. Usually some persistent excitation (PE) condition is required for the convergence of adaptive observers. However, in practice, it may happen that the PE condition is not satisfied, because the available sensor signals do not contain sufficient information for the considered recursive estimation problem, which is ill-posed. To remedy the lack of PE condition, inspired by typical methods for solving ill-posed inverse problems, this paper proposes a regularized adaptive observer for general linear time varying (LTV) systems. Regularization terms are introduced in both state and parameter estimation recursions, in order to preserve the state-parameter decoupling transformation involved in the design of the adaptive observer. Like in typical ill-posed inverse problems, regularization implies an estimation bias, which can be reduced by using prior knowledge about the unknown parameters. This work has been carried out in collaboration with Fouad Giri and Tarek Ahmed-Ali of Université de Normandie, and the results have been presented at [40].

6.2. Damage detection and localization

6.2.1. Model sensitivity clustering for damage localization

Participants: Michael Doehler, Laurent Mevel.

The purpose of this paper is the development of a working damage localization method that is applicable on real data from complex structures. To achieve this goal, robust hypothesis tests are used, the sensitivity computation of the previously published residual is revisited for more precision thanks to reduced modal truncation errors, and an adequate clustering approach is proposed for the case of a high-dimensional FE parameterization for complex structures. Finally, an application of this framework is shown for the first time on experimental data for damage localization, namely in an ambient vibration test of a 3D steel frame at the University of British Columbia. [16]

6.2.2. Robustness to temperature changes for damage localization

Participants: Laurent Mevel, Michael Doehler, Alexander Mendler.

For structures in operation, temperature has been shown to be a major nuisance to the efficiency of such methods since the modal parameters are varying not only with damage but also due to temperature variations. For detection, environmental variation is hardly taken into account in localization approaches. In this paper, we propose a sensitivity-based correction of the identified modal parameters in the damaged state with respect to the temperature field in the reference state, based on a sensitivity analysis with respect to temperature dependent parameters of the finite element model in the reference state. The approach is then applied to the Stochastic Dynamic Damage Locating Vector (SDDLV) method, where its improved performance under non-uniform temperature variations is shown in a numerical application on a beam. [18], [26]

6.2.3. Robustness to temperature changes for damage detection

Participants: Laurent Mevel, Michael Doehler, Qinghua Zhang, Eva Viefhues.

Temperature affected vibration data is evaluated with a stochastic damage detection method, which relies on a null space based residual. A new approach is proposed, using model interpolation, where a global reference model is obtained from data in the reference state at several reference temperatures. Then, for a particular testing temperature, a local reference model is derived from the global reference model. Thus, a well fitting reference null space for the formulation of the residual is available when new data is tested for damage at arbitrary temperatures. Particular attention is paid to the computation of the residual's covariance, taking into account the uncertainty related to the null space estimate. This improves the test performance, resulting in a high probability of detection (PoD) of the new interpolation approach for global and local damages compared to previous approaches. [37]

6.3. Infrared Thermography

6.3.1. Long term thermal monitoring by standard passive Infrared thermography

Participants: Thibaud Toullier, Jean Dumoulin, Laurent Mevel.

The framework of latest technological improvements in low-cost infrared cameras have brought new opportunities for long-term infrastructures monitoring. Anyway, the accurate measurement of surfaces temperatures is facing the lack of knowledge of radiatives properties of the scene. By using multi-sensors instrumentation, the measurement model can be refined to get a better estimate of the temperature. To overcome a lack of sensors instrumentation, it has been shown that online and free available climatic data can be used. [15].

6.3.2. Long term thermal monitoring by multi-spectral infrared thermography

Participants: Thibaud Toullier, Jean Dumoulin, Laurent Mevel.

Bayesian methods to estimate simultaneously the emissivity and temperature have been developed and compared to literature's methods. A radiative exchange simulator of 3D scenes have been developed to compare those different methods on numerical data. This new software uses the hardware acceleration as well as a GPGPU approach to reduce the computation time. As a consequence, obtained numerical results emphasized an advanced use of multi-spectral infrared thermography for the monitoring of structures. This simultaneous estimation enables to have an estimate of the temperature by infrared thermography with a known uncertainty. [15].

6.4. Sensor and hardware based research

6.4.1. Fiber optic and interferometry

Participants: Xavier Chapeleau, Antoine Bassil.

The assessment of Coda Wave Interferometry (CWI) and Distributed Fiber Optics Sensing (DFOS) techniques for the detection of damages in a laboratory size reinforced concrete beam is presented in this paper. The sensitivity of these two novel techniques to micro cracks is discussed and compared to standard traditional sensors. Moreover, the capacity of a DFOS technique to localize cracks and quantify crack openings is also assessed. The results show that the implementation of CWI and DFOS techniques allow the detection of early subtle changes in reinforced concrete structures until crack formation. With their ability to quantify the crack opening, following early detection and localization, DFOS techniques can achieve more effective monitoring of reinforced concrete structures. Contrary to discrete sensors, CWI and DFOS techniques cover larger areas and thus provide more efficient infrastructures asset management and maintenance operations throughout the lifetime of the structure.

6.4.2. Offset Tracking of sensor clock using Kalman filter for wireless network synchronization

Participants: David Pallier, Vincent Le Cam, Qinghua Zhang.

Wireless Sensors Networks (WSN) are more and more used in structural health monitoring applications since they represent a less expensive and non-invasive way to monitor infrastructures. Most of these applications work by merging or comparing data from several sensors located across the structure. These data often comprise measurements of physicals phenomenons evolving with time, such as acceleration and temperature. To merge or compare time-dependent data from different sensors they need to be synchronized so all the samples are time-stamped with the same time reference. An initial synchronization of the sensors is needed because sensors are independent and therefore can not be all started at the same time. Subsequent re-synchronizations are also needed since the sensors keep track of time using their imperfect local clock. This work has been presented in [34].

6.4.3. Wireless implementation of system identification techniques

Participants: Michael Doehler, Mathieu Le Pen, Vincent Le Cam, Laurent Mevel.

Embedded wireless platforms such as the PEGASE platform are appealing and suitable to collect vibration data and then perform off-line and remote computation easily. To obtain detailed modal information of large and very large structures, many sensors would be required to cover the geometry of the structure with a reasonable accuracy. However, when only a limited amount of sensors is available, large structures can be measured in several sensor setups, where some sensors remain fixed and some are moved between different measurement setups. With the sensors connected to different wireless platforms, the synchronous acquisition of data is required. In this paper, a solution of data acquisition synchronization, as well as signal processing for merging the information taking into account the change of sensor positions and environmental variability is presented.

6.4.4. Management of Cloud architectures

Participants: Jean Dumoulin, Laurent Mevel.

Cloud2IR is an autonomous software architecture, allowing multi-sensor connection, dedicated to the long term thermal monitoring of infrastructures. The system has been developed in order to cut down software integration time facilitating the system adaptation to each experiment. First, a generic unit, a data management side able to aggregate any sensor data, type or size, automatically encapsulating them in various generic data format such as hierarchical data format or cloud data such as opengis standard. This whole part is also in charge of the acquisition scenario, the local storage management and the network management. Second, a specialized unit where the sensor specific development fitted to experimental requirements are addressed. The system has been deployed on two test sites for more than one year. It aggregates various sensor data issued from infrared thermal cameras, GPS units, pyranometers, weather stations. The software and some results in outdoor conditions are discussed

MINGUS Project-Team

6. New Results

6.1. New Results

Analysis of PDEs and SPDEs

In [17], we prove the nonlinear instability of inhomogeneous steady states solutions to the Hamiltonian Mean Field (HMF) model. We first study the linear instability of this model under a simple criterion by adapting the techniques developed by the authors recently. In a second part, we extend to the inhomogeneous case some techniques developed by the authors recently and prove a nonlinear instability result under the same criterion.

In [24], we consider the non linear wave equation (NLW) on the d-dimensional torus with a smooth nonlinearity of order at least two at the origin. We prove that, for almost any mass, small and smooth solutions of high Sobolev indices are stable up to arbitrary long times with respect to the size of the initial data. To prove this result we use a normal form transformation decomposing the dynamics into low and high frequencies with weak interactions. While the low part of the dynamics can be put under classical Birkhoff normal form, the high modes evolve according to a time dependent linear Hamiltonian system. We then control the global dynamics by using polynomial growth estimates for high modes and the preservation of Sobolev norms for the low modes. Our general strategy applies to any semi-linear Hamiltonian PDEs whose linear frequencies satisfy a very general non resonance condition. The (NLW) equation on a torus is a good example since the standard Birkhoff normal form applies only when $d = 1$ while our strategy applies in any dimension.

In [20], we study semigroups generated by accretive non-selfadjoint quadratic differential operators. We give a description of the polar decomposition of the associated evolution operators as products of a selfadjoint operator and a unitary operator. The selfadjoint parts turn out to be also evolution operators generated by time-dependent real-valued quadratic forms that are studied in details. As a byproduct of this decomposition, we give a geometric description of the regularizing properties of semigroups generated by accretive non-selfadjoint quadratic operators. Finally, by using the interpolation theory, we take advantage of this smoothing effect to establish subelliptic estimates enjoyed by quadratic operators.

In [16], we prove the nonlinear orbital stability of a large class of steady states solutions to the Hamiltonian Mean Field (HMF) system with a Poisson interaction potential. These steady states are obtained as minimizers of an energy functional under one, two or infinitely many constraints. The singularity of the Poisson potential prevents from a direct run of the general strategy which was based on generalized rearrangement techniques, and which has been recently extended to the case of the usual (smooth) cosine potential. Our strategy is rather based on variational techniques. However, due to the boundedness of the space domain, our variational problems do not enjoy the usual scaling invariances which are, in general, very important in the analysis of variational problems. To replace these scaling arguments, we introduce new transformations which, although specific to our context, remain somehow in the same spirit of rearrangements tools introduced in the references above. In particular, these transformations allow for the incorporation of an arbitrary number of constraints, and yield a stability result for a large class of steady states.

In [25], we study the Boltzmann equation with external forces, not necessarily deriving from a potential, in the incompressible Navier-Stokes perturbative regime. On the torus, we establish local-in-time, for any time, Cauchy theories that are independent of the Knudsen number in Sobolev spaces. The existence is proved around a time-dependent Maxwellian that behaves like the global equilibrium both as time grows and as the Knudsen number decreases. We combine hypocoercive properties of linearized Boltzmann operators with linearization around a time-dependent Maxwellian that catches the fluctuations of the characteristics trajectories due to the presence of the force. This uniform theory is sufficiently robust to derive the incompressible Navier-Stokes-Fourier system with an external force from the Boltzmann equation. Neither smallness, nor time-decaying assumption is required for the external force, nor a gradient form, and we deal with general hard potential and cutoff Boltzmann kernels. As a by-product the latest general theories for unit Knudsen number when the force is sufficiently small and decays in time are recovered.

In [15], we show how the methods recently applied by Debussche and Weber to solve the stochastic nonlinear Schrödinger equation on $\mathbb{T}^2$ can be enhanced to yield solutions on $\mathbb{R}^2$ if the non-linearity is weak enough. We prove that the solutions remains localized on compact time intervals which allows us to apply energy methods on the full space.

In [2], we provide in this work a local in time well-posedness result for a quasilinear generalized parabolic Anderson model in dimension two $\partial_t u + \Delta\Pi(u) = g(u)\xi$. The key idea of our approach is a simple transformation of the equation which allows to treat the problem as a semilinear problem. The analysis is done within the setting of paracontrolled calculus.

In [30], we consider the Burgers equation on $H = L^2(0,1)$ perturbed by white noise and the corresponding transition semigroup $P_t D\varphi$. We prove a new formula for $PtD\varphi$ (where $\varphi : H \to \mathbb{R}$ is bounded and Borel) which depends on φ but not on its derivative. Then we deduce some consequences for the invariant measure ν of P_t as its Fomin differentiability and an integration by parts formula which generalises the classical one for gaussian measures.

In [9], we deal with the validity of a large deviation principle for the two-dimensional Navier-Stokes equation, with periodic boundary conditions, perturbed by a Gaussian random forcing. We are here interested in the regime where both the strength of the noise and its correlation are vanishing, on a length scale ε and $\delta(\varepsilon)$, respectively, with $0 < \varepsilon, \delta(\varepsilon) << 1$. Depending on the relationship between ε and $\delta(\varepsilon)$ we will prove the validity of the large deviation principle in different functional spaces.

In [30], the authors consider the transition semigroup P_t of the Φ^4_2 stochastic quantisation on the torus $\mathbb{T}^2$ and prove the following new estimate

$$|DP_t\varphi(x) \cdot h| \leq ct^{-\beta}|h|_{C^{-s}}\|\varphi\|_0(1 + |x|_{C^{-s}})^\gamma,$$

for some α, β, γ, s positive. Thanks to this estimate, we show that cylindrical functions are a core for the corresponding Kolmogorov equation. Some consequences of this fact are discussed in a final remark.

In [32], we consider a particle system with a mean-field-type interaction perturbed by some common and individual noises. When the interacting kernels are sublinear and only locally Lipschitz-continuous, relying on arguments regarding the tightness of random measures in Wasserstein spaces, we are able to construct a weak solution of the corresponding limiting SPDE. In a setup where the diffusion coefficient on the environmental noise is bounded, this weak convergence can be turned into a strong $L^p(\Omega)$ convergence and the propagation of chaos for the particle system can be established. The systems considered include perturbations of the Cucker-Smale model for collective motion.

Numerical schemes

In [7], the asymptotic behavior of the solutions of the second order linearized Vlasov-Poisson system around homogeneous equilibria is derived. It provides a fine description of some nonlinear and multidimensional phenomena such as the existence of Best frequencies. Numerical results for the $1D \times 1D$ and $2D \times 2D$ Vlasov-Poisson system illustrate the effectiveness of this approach.

In [6], we consider the problem of existence and stability of solitary traveling waves for the one dimensional discrete non linear Schrödinger equation (DNLS) with cubic nonlinearity, near the continuous limit. We construct a family of solutions close to the continuous traveling waves and prove their stability over long times. Applying a modulation method, we also show that we can describe the dynamics near these discrete traveling waves over long times.

In [4], we consider the discrete nonlinear Schrödinger equations on a one dimensional lattice of mesh h, with a cubic focusing or defocusing nonlinearity. We prove a polynomial bound on the growth of the discrete Sobolev norms, uniformly with respect to the stepsize of the grid. This bound is based on a construction of higher modified energies.

The efficient numerical solution of many kinetic models in plasma physics is impeded by the stiffness of these systems. Exponential integrators are attractive in this context as they remove the CFL condition induced by the linear part of the system, which in practice is often the most stringent stability constraint. In the literature, these schemes have been found to perform well, e.g., for drift-kinetic problems. Despite their overall efficiency and their many favorable properties, most of the commonly used exponential integrators behave rather erratically in terms of the allowed time step size in some situations. This severely limits their utility and robustness. Our goal in [29] is to explain the observed behavior and suggest exponential methods that do not suffer from the stated deficiencies. To accomplish this we study the stability of exponential integrators for a linearized problem. This analysis shows that classic exponential integrators exhibit severe deficiencies in that regard. Based on the analysis conducted we propose to use Lawson methods, which can be shown not to suffer from the same stability issues. We confirm these results and demonstrate the efficiency of Lawson methods by performing numerical simulations for both the Vlasov-Poisson system and a drift-kinetic model of a ion temperature gradient instability.

In [18], a bracket structure is proposed for the laser-plasma interaction model introduced in the physical literature, and it is proved by direct calculations that the bracket is Poisson which satisfies the Jacobi identity. Then splitting methods in time are proposed based on the Poisson structure. For the quasi- relativistic case, the Hamiltonian splitting leads to three subsystems which can be solved exactly. The conservative splitting is proposed for the fully relativistic case, and three one-dimensional conservative subsystems are obtained. Combined with the splittings in time, in phase space discretization we use the Fourier spectral and finite volume methods. It is proved that the discrete charge and discrete Poisson equation are conserved by our numerical schemes. Numerically, some numerical experiments are conducted to verify good conservations for the charge, energy and Poisson equation.

In [26], the recent advances about the construction of a Trefftz Discontinuous Galerkin (TDG) method to a class of Friedrichs systems coming from linear transport with relaxation are presented in a comprehensive setting. Application to the $2DP_N$ model are discussed, together with the derivation of new high order convergence estimates and new numerical results for the P_1 and P_3 models. More numerical results in dimension 2 illustrate the theoretical properties.

In [8], we are concerned with a formulation of Magnus and Floquet-Magnus expansions for general nonlinear differential equations. To this aim, we introduce suitable continuous variable transformations generated by operators. As an application of the simple formulas so-obtained, we explicitly compute the first terms of the Floquet-Magnus expansion for the Van der Pol oscillator and the nonlinear Schrödinger equation on the torus.

The article [11] is devoted to the construction of numerical methods which remain insensitive to the smallness of the semiclassical parameter for the linear Schrödinger equation in the semiclassical limit. We specifically analyse the convergence behavior of the first-order splitting. Our main result is a proof of uniform accuracy. We illustrate the properties of our methods with simulations.

In [10], we consider the numerical solution of highly-oscillatory Vlasov and Vlasov-Poisson equations with non-homogeneous magnetic field. Designed in the spirit of recent uniformly accurate methods, our schemes remain insensitive to the stiffness of the problem, in terms of both accuracy and computational cost. The specific difficulty (and the resulting novelty of our approach) stems from the presence of a non-periodic oscillation, which necessitates a careful ad-hoc reformulation of the equations. Our results are illustrated numerically on several examples.

In the analysis of highly-oscillatory evolution problems, it is commonly assumed that a single frequency is present and that it is either constant or, at least, bounded from below by a strictly positive constant uniformly in time. Allowing for the possibility that the frequency actually depends on time and vanishes at some instants introduces additional difficulties from both the asymptotic analysis and numerical simulation points of view. This work [13] is a first step towards the resolution of these difficulties. In particular, we show that it is still possible in this situation to infer the asymptotic behaviour of the solution at the price of more intricate computations and we derive a second order uniformly accurate numerical method.

In [12], we introduce a new methodology to design uniformly accurate methods for oscillatory evolution equations. The targeted models are envisaged in a wide spectrum of regimes, from non-stiff to highly-oscillatory. Thanks to an averaging transformation, the stiffness of the problem is softened, allowing for standard schemes to retain their usual orders of convergence. Overall, high-order numerical approximations are obtained with errors and at a cost independent of the regime.

In [1], we present an asymptotic preserving scheme based on a micro-macro decomposition for stochastic linear transport equations in kinetic and diffusive regimes. We perform a mathematical analysis and prove that the scheme is uniformly stable with respect to the mean free path of the particles in the simple telegraph model and in the general case. We present several numerical tests which validate our scheme.

In [22], a splitting strategy is introduced to approximate two-dimensional rotation motions. Unlike standard approaches based on directional splitting which usually lead to a wrong angular velocity and then to large error, the splitting studied here turns out to be exact in time. Combined with spectral methods, the so-obtained numerical method is able to capture the solution to the associated partial differential equation with a very high accuracy. A complete numerical analysis of this method is given in this work. Then, the method is used to design highly accurate time integrators for Vlasov type equations: the Vlasov-Maxwell system and the Vlasov-HMF model. Finally , several numerical illustrations and comparisons with methods from the literature are discussed.

In [23], some exact splittings are proposed for inhomogeneous quadratic differential equations including, for example, transport equations, kinetic equations, and Schrödinger type equations with a rotation term. In this work, these exact splittings are combined with pseudo-spectral methods in space to illustrate their high accuracy and efficiency.

In [14], we develop a new class of numerical schemes for collisional kinetic equations in the diffusive regime. The first step consists in reformulating the problem by decomposing the solution in the time evolution of an equilibrium state plus a perturbation. Then, the scheme combines a Monte Carlo solver for the perturbation with an Eulerian method for the equilibrium part, and is designed in such a way to be uniformly stable with respect to the diffusive scaling and to be consistent with the asymptotic diffusion equation. Moreover, since particles are only used to describe the perturbation part of the solution, the scheme becomes computationally less expensive - and is thus an asymptotically complexity diminishing scheme (ACDS) - as the solution approaches the equilibrium state due to the fact that the number of particles diminishes accordingly. This contrasts with standard methods for kinetic equations where the computational cost increases (or at least does not decrease) with the number of interactions. At the same time, the statistical error due to the Monte Carlo part of the solution decreases as the system approaches the equilibrium state: the method automatically degenerates to a solution of the macroscopic diffusion equation in the limit of infinite number of interactions. After a detailed description of the method, we perform several numerical tests and compare this new approach with classical numerical methods on various problems up to the full three dimensional case.

In [5], we revisit the old problem of compact finite difference approximations of the homogeneous Dirichlet problem in dimension 1. We design a large and natural set of schemes of arbitrary high order, and we equip this set with an algebraic structure. We give some general criteria of convergence and we apply them to obtain two new results. On the one hand, we use Padé approximant theory to construct, for each given order of consistency, the most efficient schemes and we prove their convergence. On the other hand, we use diophantine approximation theory to prove that almost all of these schemes are convergent at the same rate as the consistency order, up to some logarithmic correction.

In [28], we introduce a new Monte Carlo method for solving the Boltzmann model of rarefied gas dynamics. The method works by reformulating the original problem through a micro-macro decomposition and successively in solving a suitable equation for the perturbation from the local thermodynamic equilibrium. This equation is then discretized by using unconditionally stable exponential schemes in time which project the solution over the corresponding equilibrium state when the time step is sent to infinity. The Monte Carlo method is designed on this time integration method and it only describes the perturbation from the final state. In this way, the number of samples diminishes during the time evolution of the solution and when the final equilibrium state is reached, the number of statistical samples becomes automatically zero. The resulting method is

computationally less expensive as the solution approaches the equilibrium state as opposite to standard methods for kinetic equations which computational cost increases with the number of interactions. At the same time, the statistical error decreases as the system approaches the equilibrium state. In a last part, we show the behaviors of this new approach in comparison with standard Monte Carlo techniques and in comparison with spectral methods on different prototype problems.

In [27], we consider the three dimensional Vlasov equation with an inhomogeneous, varying direction, strong magnetic field. Whenever the magnetic field has constant intensity, the oscillations generated by the stiff term are periodic. The homogenized model is then derived and several state-of-the-art multiscale methods, in combination with the Particle-In-Cell discretisation, are proposed for solving the Vlasov-Poisson equation. Their accuracy as much as their computational cost remain essentially independent of the strength of the magnetic field. The proposed schemes thus allow large computational steps, while the full gyro-motion can be restored by a linear interpolation in time. In the linear case, extensions are introduced for general magnetic field (varying intensity and direction). Eventually, numerical experiments are exposed to illustrate the efficiency of the methods and some long-term simulations are presented.

SIMSMART Project-Team

5. New Results

5.1. Objective 1 – Rare events simulation

In [2], we present a short historical perpective of the importance splitting approach to simulate and estimate rare events, with a detailed description of several variants. We then give an account of recent theoretical results on these algorithms, including a central limit theorem for Adaptive Multilevel Splitting (AMS). Considering the asymptotic variance in the latter, the choice of the importance function, called the reaction coordinate in molecular dynamics, is also discussed. Finally, we briefly mention some worthwhile applications of AMS in various domains.

Adaptive Multilevel Splitting (AMS for short) is a generic Monte Carlo method for Markov processes that simulates rare events and estimates associated probabilities. Despite its practical efficiency, there are almost no theoretical results on the convergence of this algorithm. In [1], we prove both consistency and asymptotic normality results in a general setting. This is done by associating to the original Markov process a level-indexed process, also called a stochastic wave, and by showing that AMS can then be seen as a Fleming-Viot type particle system. This being done, we can finally apply general results on Fleming-Viot particle systems that we have recently obtained. In [1] we extend the central limit theorem to the case of synchronized branchings, where re-sampling of particles is performed after any given number of particles have been killed. The result is obtained in the generic case of Fleming-Viot particle systems.

Probability measures supported on submanifolds can be sampled by adding an extra momentum variable to the state of the system, and discretizing the associated Hamiltonian dynamics with some stochastic perturbation in the extra variable. In order to avoid biases in the invariant probability measures sampled by discretizations of these stochastically perturbed Hamiltonian dynamics, a Metropolis rejection procedure can be considered. The so-obtained scheme belongs to the class of generalized Hybrid Monte Carlo (GHMC) algorithms. In [5], we show here how to generalize to GHMC a procedure suggested by Goodman, Holmes-Cerfon and Zappa for Metropolis random walks on submanifolds, where a reverse projection check is performed to enforce the reversibility of the algorithm for large timesteps and hence avoid biases in the invariant measure. We also provide a full mathematical analysis of such procedures, as well as numerical experiments demonstrating the importance of the reverse projection check on simple toy examples.

In [24], we consider Langevin processes, which are widely used in molecular simulation to compute reaction kinetics using rare event algorithms. We prove convergence in distribution in the overdamped asymptotics. The proof relies on the classical perturbed test function (or corrector) method, which is used both to show tightness in path space, and to identify the extracted limit with a martingale problem. The result holds assuming the continuity of the gradient of the potential energy, and a mild control of the initial kinetic energy.

5.2. Objective 2 – High dimensional filtering

The work presented in [10] is about solutions for 2D multitarget tracking from image observations including its application on radar data and results on simulated data. Tracking from image observations rather than from detected points is often referred to as track–before–detect (TBD). The objective is to capture targets with a low signal–to–noise ratio (SNR) which would not be detected or tracked after data thresholding. The highly nonlinear filtering equations are approximated using a particle filter implementation. This nonlinearity emerges from the observation function which relies on the radar treatment chain, involving matched filtering of a chirp signal, and beamforming achieved from a linear phased array, leading to the raw data image. Amplitudes and phases of target–returned signals are considered as temporally fluctuating, random and unknown, creating non–deterministic contributions of the targets to the signal to deal with.

5.3. Objective 3 – Non-paramatric statistics

Production forecast errors are the main hurdle to integrate variable renewable energies into electrical power systems. Regardless of the technique, these errors are inherent in the forecast exercise, although their magnitude significantly vary depending on the method and the horizon. As power systems have to balance out these errors, their dynamic and stochastic modeling is valuable for the real time operation. The study in [23] proposes a Markov Switching Auto Regressive – MS-AR – approach. After having validated its statistical relevance, this model is used to solve the problem of the optimal management of a storage associated with a wind power plant when this virtual power plant must respect a production commitment.

5.4. Objective 4 – Model Reduction

Model reduction aims at proposing efficient algorithmic procedures for the resolution (to some reasonable accuracy) of high-dimensional systems of parametric equations. This overall objective entails many different subtasks:

1) the identification of low-dimensional surrogates of the target “solution” manifold 2) The devise of efficient methodologies of resolution exploiting low-dimensional surrogates 3) The theoretical validation of the accuracy achievable by the proposed procedures

This year, we made several contributions to these subtasks. In most of our contributions, we deviated from the standard working hypothesis involving a linear subspace surrogate.

In a first group of publications, we concentrated our attention on the so-called “sparse” low-dimensional model. In this context, we have proposed several new algorithmic solutions to decrease the computational complexity associated to projection onto this low-dimensional model. These methodologies take place in the context of “screening” procedures for LASSO. We first introduced a new screening strategy, dubbed "joint screening test", which allows the rejection of a set of atoms by performing one single test, see [4]. Our approach enables to find good compromises between complexity of implementation and effectiveness of screening. Second, we proposed two new methods to decrease the computational cost inherent to the construction of the (so-called) "safe region". Our numerical experiments show that the proposed procedures lead to significant computational gains as compared to standard methodologies, see [11]. We finally showed in another work that the main concepts underlying screening procedures can be extended to different families of convex optimization problems, see [22].

Another avenue of research has been the study of the sparse surrogate in the context of “continuous” dictionaries, where the elementary signals forming the decomposition catalog are functions of some parameters taking its values in some continuously-valued domain. In this context, we contributed to the theoretical characterization of the performance of some well-known algorithmic procedure, namely “orthogonal matching pursuit” (OMP). More specifically, we proposed the first theoretical analysis of the behavior of OMP in the continuous setup, see [12], [17], [21]. We also provided a new connection between two popular low-rank approximations of continuous dictionaries, namely the “polar” and “SVD” approximations, see [9].

The tools exploited in the field of model-order reduction and sparsity have found some particular applicative field in geophysics and fluid mechanics. In [8], [7], we derived procedures based on sparse representations to localize the positions of particles in a moving fluid. In [16], [15], [14], [26], [27], we designed learning methodologies to learn the dynamical model underlying a set of observed data.

5.5. Miscellaneous

In [25], we devise methods of variance reduction for the Monte Carlo estimation of an expectation of the type $\mathbb{E}[\phi(X,Y)]$, when the distribution of X is exactly known. The key general idea is to give each individual of a sample a weight, so that the resulting weighted empirical distribution has a marginal with respect to the variable X as close as possible to its target. We prove several theoretical results on the method, identifying settings where the variance reduction is guaranteed. We perform numerical tests comparing the methods and demonstrating their efficiency.

In [6], we consider the problem of predicting a categorical variable based on groups of inputs. Some methods have already been proposed to elaborate classifi- cation rules based on groups of variables (e.g. group lasso for logistic regression). However, to our knowledge, no tree-based approach has been proposed to tackle this issue. Here, we propose the Tree Penalized Linear Discriminant Analysis algorithm (TPLDA), a new-tree based approach which constructs a classification rule based on groups of variables.

DYLISS Project-Team

7. New Results

7.1. Scalable methods to query data heterogenity

Participants: Emmanuelle Becker, Lucas Bourneuf, Olivier Dameron, Xavier Garnier, Vijay Ingalalli, Marine Louarn, Yann Rivault, Anne Siegel.

Increasing life science resources re-usability using Semantic Web technologies [*E. Becker, O. Dameron, X. Garnier, V. Ingallali, M. Louarn, Y. Rivault, A. Siegel*] [25], [18], [29], [31], [23], [27], [28]. Our work was focused on assessing to what extent Semantic Web technologies also facilitate reproducibility and reuse of life sciences studies involving pipelines that compute associations between entities according to intermediary relations and dependencies.

- We followed on 2018 action exploratoire Inria by studying possible optimizations for federated SPARQL queries [31]
- We considered a case-study in systems biology ([Regulatorycircuits link]), which provides tissue-specific regulatory interaction networks to elucidate perturbations across complex diseases. We relied on this structure and used Semantic Web technologies (i) to integrate the Regulatory Circuits data, and (ii) to formalize the analysis pipeline as SPARQL queries. Our result was a 335,429,988 triples dataset on which two SPARQL queries were sufficient to extract each single tissue-specific regulatory network.
- A second case-study concerned public health data for reusing electronic health data, selecting patients, identifying specific events and interpreting results typically requires biomedical knowledge [64]. We developed the queryMed R package [18], [29]. It aims to facilitate the integration of medical and pharmacological knowledge stored in formats compliant with the Linked Data paradigm (e.g. OWL ontologies and RDF datasets) into the R statistical programming environment. We showed that linking a medical database of 1003 critical limb ischemia (CLI) patients to ontologies allowed us to identify all the drugs prescribed for CLI and also to detect one contraindicated prescription for one patient. We also investigated temporal models of care sequences for the exploration of medico-administrative data as part of Johanne Bakalara's PhD, supervized with Thomas Guyet (Lacodam) and Emmanuel Oger (Repères).
- We pursued the development of AskOmics [27]. Version 3 adds the capability to generate the graph of entity types (aka abstraction) from typed RDF datasets, improved management of entity hierarchies and support for federated queries on external SPARQL endpoints such as UniProt and neXtProt.

Graph compression and analysis [*L. Bourneuf*] [26], [24]. Because of the increasing size and complexity of available graph structures in experimental sciences like molecular biology, techniques of graph visualization tend to reach their limit.

- We developed the Biseau approach, a programming environment aiming at simplifying the visualization task. Biseau takes advantage of Answer Set Programming and shows as a use-case how Formal Concept Analysis can be efficiently described at the level of its properties, without needing a costly development process. It reproduces the core results of existing tools like LatViz or In-Close.
- We formalized a graph compression search space in order to provide approximate solutions to the NP-complete problem of computing a lossless compression of the graph based on the search of cliques and bicliques. Our conclusion is that the search for graph compression can be usefully associated with the search for patterns in a concept lattice and that, conversely, confusing sets of objects and attributes brings new interesting problems for FCA.

7.2. Metabolism: from enzyme sequences to systems ecology

Participants: Méziane Aite, Arnaud Belcour, Mael Conan, François Coste, Clémence Frioux, Jeanne Got, Anne Siegel, Hugo Talibart.

Modelling proteins with long distance dependencies [*F. Coste, H. Talibart*] [15], [30], [30]

- We proposed to use information on protein contacts to train probabilistic context-free grammars representing families of protein sequences. We developed the theory behind the introduction of contact constraints in maximum-likelihood and contrastive estimation schemes and implemented it in a machine learning framework for protein grammars. Evaluation showed high fidelity of grammatical descriptors to protein structures, improved precision in recognizing sequences and the ability to model a meta-family of proteins that could not be modeled by classical approaches [15].
- We then investigated the problem of modeling proteins with crossing dependencies. Motivated by their success on contact prediction, we propose to use Potts models for the purposes of modeling proteins and searching. We developed ComPotts a tool for optimal alignment and comparison of Potts models, enabling to take into account the coevolution of residues for the search of protein homologs [30], [40].

Large-scale eukaryotic metabolic network reconstruction [*A. Siegel, C. Frioux, M. Aite, A. Belcour, J. Got, N. Théret, M. Conan*] [17], [14], [38]. Metabolic network reconstruction has attained high standards but is still challenging for complex organisms such as eukaryotes.

- *Large-scale eukaryotic metabolic network reconstruction*: We participated to the reconstruction of a genome–scale metabolic network for the brown Algae *Saccharina japonica* and *Cladosiphon okamuranus* in order to shed light of the specificities on the carotenoid biosynthesis Pathway.
- *Metabolic pathway inference from non genomic data*: We designed methods for the identification of metabolic pathways for which enzyme information is not precise enough. As an application study, we focused on Heterocyclic Aromatic Amines (HAAs), which are environmental and food contaminants classified as probable carcinogens. Our approach based on a refinement of molecular predictions with enzyme activity scores allows to accurately predict HAAs biotransformation and their potentials DNA reactive compounds [54].

Systems ecology: design of microbial consortia [*C. Frioux, A. Belcour, J. Got, M. Aite, A. Siegel*] [21], [22], [34], [33].

- We participated to the application of our methods to algal-microbial consortia, with good preliminary results, and presented them as an invited conference [22].

7.3. Regulation and signaling: detecting complex and discriminant signatures of phenotypes

Participants: Catherine Belleannée, Célia Biane-Fourati, Samuel Blanquart, Olivier Dameron, Maxime Folschette, Nicolas Guillaudeux, Marine Louarn, François Moreews, Anne Siegel, Nathalie Théret, Pierre Vignet, Méline Wéry.

Creation of predictive functional signaling networks [*M. Folschette, N. Théret*] [16].

- Integrating genome-wide gene expression patient profiles with regulatory knowledge is a challenging task because of the inherent heterogeneity, noise and incompleteness of biological data. We proposed an automatic pipeline to extract automatically regulatory knowledge from pathway databases and generate novel computational predictions related to the state of expression or activity of biological molecules. We applied it in the context of hepatocellular carcinoma (HCC) progression, and evaluated the precision and the stability of these computational predictions. Our computational model predicted the shifts of expression of 146 initially non-observed biological components. Our predictions were validated at 88% using a larger experimental dataset and cross-validation techniques.

Experimental evidences of transcript predictions [*C. Belleannée, S. Blanquart, N. Guillaudeux*] [13].

- We designed comparative-genomics based models of gene structures through genes comparisons across species. These models enable to predict putative transcript isoforms in a species given the knowledge available in other species [46]. We recently published a first experimental validation of such a predicted transcriptome [13]. In this work, transcript isoforms of the human TRPM8 gene yield transcript predictions in the mouse TRPM8 gene, which are experimentally validated using targeted PCR in mouse tissues. This work also provides a first attempt to estimate origin of new isoforms during the gene evolution.
- In another collaboration with IGDR, we considered a multi-species gene comparison including human, mouse and dog [39]. This work reveals global trends of the gene isoform sets evolution, suggesting a extremely high plasticity of alternative transcription and alternative splicing propensities in those three species. This work moreover provides experimental evidences of the predicted transcripts based on public RNAseq data, highlighting the tissue specificity of isoform expression across species.

Formalizing and enriching phenotype signatures using Boolean networks [*C. Biane-Fourati, M. Wéry, A. Siegel, O. Dameron*] [20], [42], [22]

- We used Formal Concept Analysis as a symbolic bi-clustering techniques to classify and sort the steady states of a Boolean network according to biological signatures based on the hierarchy of the roles the network components play in the phenotypes. We applied our approach on a T helper lymphocyte (Th) differentiation network with a set of signatures corresponding to the sub-types of Th. This led to the identification and prediction of a new hybrid sub-type later confirmed by the literature.

EMPENN Project-Team

7. New Results

7.1. Research axis 1: Medical Image Computing in Neuroimaging

Extraction and exploitation of complex imaging biomarkers involve an imaging processing workflow that can be quite complex. This goes from image physics and image acquisition, image processing for quality control and enhancement, image analysis for features extraction and image fusion up to the final application which intends to demonstrate the capability of the image processing workflow to issue sensitive and specific markers of a given pathology. In this context, our objectives in the recent period were directed toward following major methodological topics:

7.1.1. Diffusion imaging

7.1.1.1. Free water estimation using single-shell diffusion-weighted images

Participant: Emmanuel Caruyer.

Free-water estimation requires the fitting of a bi-compartment model, which is an ill-posed problem when using only single-shell data. Its solution requires optimization, which relies on an initialization step. We propose a novel initialization approach, called "Freewater EstimatoR using iNtErpolated iniTialization" (FERNET), which improves the estimation of free water in edematous and infiltrated peritumoral regions, using single-shell diffusion MRI data. The method has been extensively investigated on simulated data and healthy and brain tumor datasets, demonstrating its applicability on clinically acquired data. Additionally, it has been applied to data from brain tumor patients to demonstrate the improvement in tractography in the peritumoral region [57].

7.1.1.2. Multi-dimensional diffusion MRI sampling scheme: B-tensor design and accurate signal reconstruction

Participant: Emmanuel Caruyer.

B-tensor encoding enables the separation of isotropic and anisotropic tensors. However, little consideration has been given as to how to design a B-tensor encoding sampling scheme. In this work, we propose the first 4D basis for representing the diffusion signal acquired with B-tensor encoding. We study the properties of the diffusion signal in this basis to give recommendations for optimally sampling the space of axisymmetric b-tensors. We show, using simulations, that the proposed sampling scheme enables accurate reconstruction of the diffusion signal by expansion in this basis using a clinically feasible number of samples [24].

This work was done in collaboration with A. Bates, Australian National University and Al. Daducci, University of Verona.

7.1.1.3. Optimal selection of diffusion-weighting gradient waveforms using compressed sensing and dictionary learning

Participants: Raphaël Truffet, Emmanuel Caruyer, Christian Barillot.

Acquisition sequences in diffusion MRI rely on the use of time-dependent magnetic field gradients. Each gradient waveform encodes a diffusion-weighted measure; a large number of such measurements are necessary for the in vivo reconstruction of microstructure parameters. We propose here a method to select only a subset of the measurements, while being able to predict the unseen data using compressed sensing. We learn a dictionary using a training dataset generated with Monte-Carlo simulations; we then compare two different heuristics to select the measures to use for the prediction. We found that an undersampling strategy limiting the redundancy of the measures allows for a more accurate reconstruction when compared with random undersampling with similar sampling rate [49].

7.1.1.4. Geometric evaluation of distortion correction methods in diffusion MRI of the spinal cord

Participants: Haykel Snoussi, Emmanuel Caruyer, Olivier Commowick, Benoit Combès, Élise Bannier, Christian Barillot.

Acquiring and processing Diffusion MRI in spinal cord present inherent challenges. Differences in magnetic susceptibility between soft tissues, air and bones make the magnetic field non uniform in spinal cord. In this context, various procedures were proposed for correcting inhomogeneity-induced distortions; in this work, we propose novel geometric statistics to measure the alignment of the reconstructed diffusion model with the apparent centerline of the spine. In parallel of the correlation with an anatomical T2-weighted image, we show the utility of these statistics to study and evaluate the impact of distortion correction by comparing three correction methods using a pair of images acquired with reversed gradient polarity [48].

This work was done in collaboration with Anne Kerbrat, Neuropoly Montréal and Julien Cohen-Adad from NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada.

7.1.2. Arterial Spin Labeling

7.1.2.1. Acquisition duration in resting-state arterial spin labeling. How long is enough?
Participants: Corentin Vallée, Pierre Maurel, Isabelle Corouge, Christian Barillot.

Resting-state Arterial Spin Labeling (rs-ASL) is a rather confidential method compared to resting-state BOLD but it drives great prospects with respect to potential clinical applications. By enabling the study of cerebral blood flow maps, rs-ASL can lead to significant clinical subject-scaled applications as CBF is a biomarker in neuropathology. An important parameter to consider in functional imaging is the acquisition duration. Despite directly impacting practicability and functional networks representation, there is no standard for rs-ASL. Our work here focuses on strengthening the confidence in ASL as a rs-fMRI method, and on studying the influence of the acquisition duration. To this end, we acquired a long rs-ASL sequence and assessed the quality of typical functional brain networks quality over time compared to gold-standard networks. Our results show that after 14min of duration acquisition, functional networks representation can be considered as stable [58], [50].

7.1.2.2. Patch-based super-resolution of arterial spin labeling magnetic resonance images
Participants: Cédric Meurée, Pierre Maurel, Jean-Christophe Ferré, Christian Barillot.

Arterial spin labeling is a magnetic resonance perfusion imaging technique that, while providing results comparable to methods currently considered as more standard concerning the quantification of the cerebral blood flow, is subject to limitations related to its low signal-to-noise ratio and low resolution. In this work, we investigated the relevance of using a non-local patch-based super-resolution method driven by a high-resolution structural image to increase the level of details in arterial spin labeling images. This method was evaluated by comparison with other image resolution increasing techniques on a simulated dataset, on images of healthy subjects and on images of subjects diagnosed with brain tumors, who had a dynamic susceptibility contrast acquisition. The influence of an increase of ASL images resolution on partial volume effects was also investigated in this work [16].

The development of this super-resolution algorithm in the context of the PhD of Cédric Meurée founded by Siemens Healthineers conducted to a stay of one month of the PhD candidate in Erlangen, during summer 2018. This immersion into the neuro-development team allowed him to integrate the proposed solution with tools in use within this team. Part of the work also consisted in reducing the computation, a factor of 5 being achieved at the end of these four weeks.

7.1.3. Atlases

7.1.3.1. Unbiased longitudinal brain atlas creation using robust linear registration and log-Euclidean framework for diffeomorphisms
Participants: Antoine Legouhy, Olivier Commowick, Christian Barillot.

We have defined a new method to create a diffeomorphic longitudinal (4D) atlas composed of a set of 3D atlases each representing an average model at a given age. This is achieved by generalizing atlasing methods to produce atlases unbiased with respect to the initial reference up to a rigid transformation and ensuring diffeomorphic deformations thanks to the Baker-Campbell-Hausdorff formula and the log-Euclidean framework for diffeomorphisms. Subjects are additionally weighted using an asymmetric function to closely match specified target ages. Creating a longitudinal atlas also implies dealing with subjects with large brain differences that can lead to registration errors. This is overcome by a robust rigid registration based on polar decomposition. We illustrated these techniques for the creation of a 4D pediatric atlas, showing their ability to create a temporally consistent atlas [22].

This work was done in collaboration with François Rousseau, IMT Atlantique, LaTIM U1101 INSERM, Brest, France, under the ANR MAIA project.

7.1.3.2. Online atlasing using an iterative centroid

Participants: Antoine Legouhy, Olivier Commowick, Christian Barillot.

Online atlasing, i.e. incrementing an atlas with new images as they are acquired, is key when performing studies on databases very large or still being gathered. We proposed to this end a new diffeomorphic online atlasing method without having to perform again the atlasing process from scratch. New subjects are integrated following an iterative procedure gradually shifting the centroid of the images to its final position, making it computationally cheap to update regularly an atlas as new images are acquired (only needing a number of registrations equal to the number of new subjects). We evaluated this iterative centroid approach through the analysis of the sharpness and variance of the resulting atlases, and the transformations of images, comparing their deviations from a conventional method using Guimond's method. We demonstrated that the transformations divergence between the two approaches is small and stable, and that both atlases reach equivalent levels of image quality [42].

This work was done in collaboration with François Rousseau, IMT Atlantique, LaTIM U1101 INSERM, Brest, France, under the ANR MAIA project.

7.1.4. Neurofeedback

7.1.4.1. Learning bi-modal EEG-fMRI neurofeedback to improve neurofeedback in EEG only

Participants: Claire Cury, Pierre Maurel, Giulia Lioi, Christian Barillot.

In neurofeedback (NF), a new kind of data are available: electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) acquired simultaneously during bi-modal EEG-fMRI neurofeedback. These two complementary techniques have only recently been integrated in the context of NF for brain rehabilitation protocols. Bi-modal NF (NF-EEG-fMRI) combines information coming from two modalities sensitive to different aspect of brain activity, therefore providing a higher NF quality. However, the use of the MRI scanner is cumbersome and exhausting for patients. We presented, a novel methodological development, able to reduce the use of fMRI while providing to subjects NF-EEG sessions of quality comparable to the bi-modal NF sessions. We proposed an original alternative to the ill-posed problem of source reconstruction. We designed a non-linear model considering different frequency bands, electrodes and temporal delays, with a structured sparse regularisation. Results show that our model is able to significantly improve the quality of NF sessions over what EEG could provide alone. We tested our method on 17 subjects that performed three NF-EEG-fMRI sessions each [30].

7.1.4.2. Can we learn from coupling EEG-fMRI to enhance neuro-feedback in EEG only?

Participants: Claire Cury, Pierre Maurel, Christian Barillot.

Neurofeedback (NF) measures brain activation during a task, and gives back to the subject a score reflecting his/her performance that he/she tries to improve. Among noninvasive functional brain imaging modalities, the most used in NF, are electro-encephalography (EEG) and the functional magnetic resonance imaging (fMRI). EEG measures the electrical activity of the brain through channels located on the scalp, with an excellent temporal resolution (milliseconds), but has a limited spatial resolution due to the well-known ill-posed inverse problem of source reconstruction. Also NF-EEG (NF session with NF scores extracted from EEG) is not easy to control since it comes from mixtures of propagating electric potential fluctuations. Blood oxygenation level dependent (BOLD) fMRI measures neuro-vascular activity, easier to control, with an excellent spatial resolution, making NF-fMRI (NF session with NF scores extracted from BOLD-fMRI) an adequate modality for NF. However its temporal resolution is only of a few seconds, and it is a costly, exhausting for subjects and time consuming modality. Since those modalities are complementary, their combined acquisition is actively investigated, as well as the methodology to extract information from fMRI with EEG which is the easiest modality to use [Abreu et al. 2018]. Our challenge is to learn EEG activation patterns from NF-fMRI scores extracted during a NF session using coupled EEG-fMRI data (NF-EEG-fMRI) to improve NF scores when using EEG only [29].

7.1.5. Deep learning

7.1.5.1. Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from MRI data

Participants: Antoine Ackaouy, Olivier Commowick, Christian Barillot, Francesca Galassi.

Automatic segmentation of Multiple Sclerosis (MS) lesions from Magnetic Resonance Imaging (MRI) images is essential for clinical assessment and treatment planning of MS. Recent years have seen an increasing use of Convolutional Neural Networks (CNNs) for this task. Although these methods provide accurate segmentation, their applicability in clinical settings remains limited due to a reproducibility issue across different image domains. MS images can have highly variable characteristics across patients, MRI scanners and imaging protocols. Retraining a supervised model with data from each new domain is not a feasible solution because it requires manual annotation from expert radiologists. In this work, we explored an unsupervised solution to the problem of domain shift. We presented a framework, Seg-JDOT, which adapts a deep model so that samples from a source domain and samples from a target domain sharing similar representations will be similarly segmented. We evaluated the framework on a multi-site dataset, MICCAI 2016, and showed that the adaptation towards a target site can bring remarkable improvements in a model performance over standard training [54].

This work was done in collaboration with Nicolas Courty, Obelix team, IRISA laboratory from University of Bretagne Sud.

7.1.5.2. Deep learning for multi-site MS lesions segmentation: two-step intensity standardization and generalized loss function.

Participants: Francesca Galassi, Olivier Commowick, Christian Barillot.

We presented an improved CNN framework for the segmentation of Multiple Sclerosis (MS) lesions from multi-modal MRI. It uses a two-step intensity normalization and a cascaded network with cost sensitive learning. Performance was assessed on a public multi-site data-set [35].

7.2. Research axis 2: Applications in Neuroradiology and Neurological Disorders

Our objectives is also to provide new computational solutions for our target clinical applications (radiology, neurology, psychiatry and rehabilitation...), allowing a more appropriate representation of data for image analysis and the detection of biomarkers specific to a form or grade of pathology, or specific to a population of subjects. In this section, we present our contributions in different clinical applications.

7.2.1. Rehabilitation

7.2.1.1. Efficacy of EEG-fMRI Neurofeedback for stroke rehabilitation in relation to the DTI structural damage: a pilot study.
Participants: Giulia Lioi, Mathis Fleury, Christian Barillot, Isabelle Bonan.

Recent studies have shown the potential of neurofeedback (NF) for motor rehabilitation after stroke. The majority of these NF approaches have relied solely on one imaging technique: mostly on EEG recordings. Recent study have gone further, revealing the potential of integrating complementary techniques such as EEG and fMRI to achieve a more specific regulation. In this exploratory work, multisession bimodal EEG-fMRI NF for upper limb motor recovery was tested in four stroke patients.The feasibility of the NF training was investigated with respect to the integrity of the corticospinal tract (CST), a well-established predictor of the potential for clinical improvement. Results indicated that patients exhibiting a high degree of integrity of the ipsilesional CST showed significant increased activation of the ipsilesional M1 at the end of the training ($p<<0.001$, Wilcoxon test). These preliminary findings confirm the critical role of the CST integrity for stroke motor recovery and indicate that this is importantly related also to functional brain regulation of the ipsilesional motor cortex [43].

7.2.2. Multiple sclerosis

7.2.2.1. Tissue microstructure information from T2 relaxometry and diffusion MRI can identify multiple sclerosis (MS) lesions undergoing blood-brain barrier breakdown (BBB)
Participants: Olivier Commowick, Christian Barillot.

Gadolinium-based contrast agents (GBCA) play a critical role in identifying MS lesions undergoing BBB which is of high clinical importance. However, repeated use of GBCAs over a long period of time and the risks associated with administering it to patients with renal complications has mandated for greater caution in its usage. In this work we explored the plausibility of identifying MS lesions undergoing BBB from tissue microstructure information obtained from T2 relaxometry and dMRI data. We also proposed a framework to predict MS lesions undergoing BBB using the tissue microstructure information and demonstrated its potential on a test case [26].

7.2.2.2. Neural basis of irony in patients with Multiple Sclerosis: an exploratory fMRI study
Participants: Quentin Duché, Élise Bannier.

Irony is a form of non-literal language that is characterized by the opposition between the literal meaning of a statement and the message that the speaker wishes to convey. Knowledge about the neural bases of non-literal language has largely developed in recent years from injury studies or more recently through data from functional imaging studies. Multiple sclerosis (MS) is a neurodegenerative disease that, in addition to cognitive dysfunction, results in variable impairment of theory of mind and non-literal language skills. This work aims at exploring neural basis underpinning the comprehension of irony in MS patients compared to a group of healthy subjects. The results suggest that multiple sclerosis patients require higher left hemisphere resources than healthy controls to understand irony [32].

This work is done in collaboration with by Florian Chapelain (Pôle Saint Hélier), Philippe Gallien (Pôle Saint Hélier) and Virginie Dardier (Université Rennes 2).

7.2.2.3. Joint assessment of brain and spinal cord motor tract damage in patients with early relapsing remitting multiple sclerosis (RRMS): predominant impact of spinal cord lesions on motor function
Participants: Benoit Combès, Élise Bannier, Haykel Snoussi, Jean-Christophe Ferré, Christian Barillot.

The effect of structural multiple sclerosis damage to the corticospinal tract (CST) has been separately evaluated in the brain and spinal cord (SC), even though a cumulative impact is suspected. In this work, we evaluated CST damages on both the cortex and cervical SC, and examine their relative associations with motor function, measured both clinically and by electrophysiology. This study highlights the major contribution of SC lesions to CST damage and motor function abnormalities [8].

This work was done in collaboration with Anne Kerbrat (Neuropoly Montréal) and Raphael Chouteau (CHU Rennes).

7.2.2.4. Spatial distribution of multiple sclerosis lesions in the cervical spinal cord

Participants: Élise Bannier, Gilles Edan.

Spinal cord lesions detected on MRI hold important diagnostic and prognostic value for multiple sclerosis. Our aim was to explore the spatial distribution of multiple sclerosis lesions in the cervical spinal cord, with respect to clinical status. We included 642 suspected or confirmed multiple sclerosis patients (31 clinically isolated syndrome, and 416 relapsing-remitting, 84 secondary progressive, and 73 primary progressive multiple sclerosis) from 13 clinical sites. With an automatic publicly-available analysis pipeline we produced voxelwise lesion frequency maps to identify predilection sites in various patient groups characterized by clinical subtype, Expanded Disability Status Scale score and disease duration. We also measured absolute and normalized lesion volumes in several regions of interest using an atlas-based approach, and evaluated differences within and between groups. The lateral funiculi were more frequently affected by lesions in progressive subtypes than in relapsing in voxelwise analysis ($P < 0.001$), which was further confirmed by absolute and normalized lesion volumes ($P < 0.01$). The central cord area was more often affected by lesions in primary progressive than relapse-remitting patients ($P < 0.001$). Between white and grey matter, the absolute lesion volume in the white matter was greater than in the grey matter in all phenotypes ($P < 0.001$); however when normalizing by each region, normalized lesion volumes were comparable between white and grey matter in primary progressive patients. Lesions appearing in the lateral funiculi and central cord area were significantly correlated with Expanded Disability Status Scale score ($P < 0.001$). High lesion frequencies were observed in patients with a more aggressive disease course, rather than long disease duration. Lesions located in the lateral funiculi and central cord area of the cervical spine may influence clinical status in multiple sclerosis. This work shows the added value of cervical spine lesions, and provides an avenue for evaluating the distribution of spinal cord lesions in various patient groups [14].

This work was done in collaboration with Julien Cohen-Adad (Neuropoly, Montreal) and Anne Kerbrat (Neuropoly Montréal).

7.2.2.5. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks

Participants: Élise Bannier, Gilles Edan.

The goal of this study was to develop a fully-automatic framework - robust to variability in both image parameters and clinical condition - for segmentation of the spinal cord and intramedullary MS lesions from conventional MRI data of MS and non-MS cases. Scans of 1042 subjects (459 healthy controls, 471 MS patients, and 112 with other spinal pathologies) were included in this multi-site study (n=30). Data spanned three contrasts (T1-, T2-, and $T2_*$-weighted) for a total of 1943vol and featured large heterogeneity in terms of resolution, orientation, coverage, and clinical conditions. The proposed cord and lesion automatic segmentation approach is based on a sequence of two Convolutional Neural Networks (CNNs). CNNs were trained independently with the Dice loss. When compared against manual segmentation, our CNN-based approach showed a median Dice of 95% vs. 88% for PropSeg ($p \leq 0.05$), a state-of-the-art spinal cord segmentation method. Regarding lesion segmentation on MS data, our framework provided a Dice of 60%, a relative volume difference of -15%, and a lesion-wise detection sensitivity and precision of 83% and 77%, respectively. In this study, we introduce a robust method to segment the spinal cord and intramedullary MS lesions on a variety of MRI contrasts. The proposed framework is open-source and readily available in the Spinal Cord Toolbox.

This work was done in collaboration with Julien Cohen-Adad (Neuropoly, Montreal) and Anne Kerbrat (Neuropoly Montréal).

7.2.3. Arterial Spin Labeling in pediatric populations

7.2.3.1. Changes in brain perfusion in successive arterial spin labeling MRI scans in neonates with hypoxic-ischemic encephalopathy

Participants: Maia Proisy, Isabelle Corouge, Antoine Legouhy, Christian Barillot, Jean-Christophe Ferré.

The primary objective of this study was to evaluate changes in cerebral blood flow (CBF) using arterial spin labeling MRI between day 4 of life (DOL4) and day 11 of life (DOL11) in neonates with hypoxic-ischemic encephalopathy (HIE) treated with hypothermia. The secondary objectives were to compare CBF values between the different regions of interest (ROIs) and between infants with ischemic lesions on MRI and infants with normal MRI findings. We prospectively included all consecutive neonates with HIE admitted to the neonatal intensive care unit of our institution who were eligible for therapeutic hypothermia. Each neonate systematically underwent two MRI examinations as close as possible to day 4 (early MRI) and day 11 (late MRI) of life. We proposed an innovative processing pipeline for morphological and ASL data suited to neonates that enable automated segmentation to obtain CBF values over ROIs. We evaluated CBF on two successive scanswithin thefirst 15 days of life in the same subjects. ASL imaging in asphyxiated neonates seems more relevant when used relatively early, in the first days of life. The correlation of intra-subject changes in cerebral perfusion between early and late MRI with neurodevelopmental outcome warrants investigation in a larger cohort, to determine whether the CBF pattern change can provide prognostic information beyond that provided by visible structural abnormalities on conventional MRI [18], [47].

7.2.4. Cerebral blood flow in sickle cell populations

7.2.4.1. White matter has impaired resting oxygen delivery in sickle cell patients

Participant: Julie Coloigner.

Although modern medical management has lowered overt stroke occurrence in patients with sickle cell disease (SCD), progressive white matter (WM) damage remains common. It is known that cerebral blood flow (CBF) increases to compensate for anemia, but sufficiency of cerebral oxygen delivery, especially in the WM, has not been systematically investigated. Cerebral perfusion was measured by arterial spin labeling in 32 SCD patients (age range: 10-42 years old, 14 males, 7 with hemoglobin SC, 25 hemoglobin SS) and 25 age and race-matched healthy controls (age range: 15-45 years old, 10 males, 12 with hemoglobin AS, 13 hemoglobin AA); 8/24 SCD patients were receiving regular blood transfusions and 14/24 non-transfused SCD patients were taking hydroxyurea. Imaging data from control subjects were used to calculate maps for CBF and oxygen delivery in SCD patients and their T-score maps. Whole brain CBF was increased in SCD patients with a mean T-score of 0.5 and correlated with lactate dehydrogenase ($r2 = 0.58$, $P < 0.0001$). When corrected for oxygen content and arterial saturation, whole brain and gray matter (GM) oxygen delivery were normal in SCD, but WM oxygen delivery was 35% lower than in controls. Age and hematocrit were the strongest predictors for WM CBF and oxygen delivery in patients with SCD. There was spatial co-localization between regions of low oxygen delivery and WM hyperintensities on T2 FLAIR imaging. To conclude, oxygen delivery is preserved in the GM of SCD patients, but is decreased throughout the WM, particularly in areas prone to WM silent strokes [7].

This work was done in collaboration with Natasha Leporé and her team, Children's hospital Los Angeles, University of Southern California, USA.

7.2.5. Alzheimer disease

7.2.5.1. Abnormal fMRI response in sub-hippocampal structures: how prior knowledge impairs memory in AD

Participants: Quentin Duché, Pierre-Yves Jonin.

Early Alzheimer's disease typically impairs associative learning abilities, up to 18 years before dementia. Importantly, patients' concerns refer to their daily routine, meaning that they lack associative memory for highly familiar stimuli. However, most of the tests involve much less familiar stimuli (e.g. isolated words). It follows that we ignore whether prior knowledge about memoranda alters memory formation and its neural correlates in Alzheimer's Disease. Here, we aimed at manipulating prior knowledge available at encoding and repetition to investigate whether prior knowledge could alter the neural underpinnings of associative encoding, in a way sensitive to early AD. The results suggest that distinct forms of prior knowledge may drive partly non-overlapping brain networks at encoding, and in turn these regions differentially contribute to successful memory formation. Thus, our finding that sub-hippocampal, not hippocampal, activation underlie the inability of the patients to benefit remote prior knowledge in new learning opens perspectives for further diagnostic and prognostic markers development [37].

7.2.5.2. Learning what you know: how prior knowledge impairs new associative learning in early AD.

Participants: Pierre-Yves Jonin, Quentin Duché, Élise Bannier, Isabelle Corouge, Jean-Christophe Ferré, Christian Barillot.

While associative memory impairment is a core feature of prodromal Alzheimer's Disease (AD), whether prior knowledge affects associative learning is largely overlooked. Stimuli repetition yields suppression or enhancement of the BOLD signal, allowing the functional mapping of brain networks. We addressed the role of prior knowledge in associative encoding by manipulating repetition and familiarity of the memoranda in a subsequent memory fMRI study design. 17 patients with prodromal AD (AD-MCI) and 19 Controls learned face-scene associations presented twice in the scanner. Pre-experimental knowledge trials (PEK) involved famous faces while in Experimental Knowledge trials (EK), unknown faces familiarized before scanning were used. Study events were sorted as associative hits, associative misses or misses after a recognition test outside the scanner. We computed the Repetition X Prior knowledge interaction contrast to test whether the encoding networks differed along with prior knowledge, then looked for subsequent associative memory effects in the resulting clusters. PEK and EK yielded similar associative memory performance in AD-MCI, while PEK increased associative memory by 28% in Controls. Repetition effects were modulated by Prior knowledge in Controls, but AD-MCI showed aberrant repetition effects. Subsequent memory effects were observed only in Controls for PEK in the right subhippocampal structures. By contrast, in both groups, EK triggered a subsequent memory effect in the right hippocampus. Provided that tau pathology starts within anterior subhippocampal regions in early AD, our findings that subhippocampal, not hippocampal, involvement underlies the inability of the patients to benefit from PEK open innovative clinical and research perspectives [38].

7.2.6. Depression

7.2.6.1. White matter abnormalities in depression: a categorical and phenotypic diffusion MRI study.

Participants: Julie Coloigner, Olivier Commowick, Isabelle Corouge, Christian Barillot.

Mood depressive disorder is one of the most disabling chronic diseases with a high rate of everyday life disability that affects 350 million people around the world. Recent advances in neuroimaging have reported widespread structural abnormalities, suggesting a dysfunctional frontal-limbic circuit involved in the pathophysiological mechanisms of depression. However, a variety of different white matter regions has been highlighted and these results lack reproducibility of such categorical-based biomarkers. These inconsistent results might be attributed to various factors: actual categorical definition of depression as well as clinical phenotype variability. In this study, we 1/ examined WM changes in a large cohort (114 patients) compared to a healthy control group and 2/ sought to identify specific WM alterations in relation to specific depressive phenotypes such as anhedonia (i.e. lack of pleasure), anxiety and psychomotor retardation –three core symptoms involved in depression. Consistent with previous studies, reduced white matter was observed in the genu of the corpus callosum extending to the inferior fasciculus and posterior thalamic radiation, confirming a frontal-limbic circuit abnormality. Our analysis also reported other patterns of increased fractional anisotropy and axial diffusivity as well as decreased apparent diffusion coefficient and radial diffusivity in the splenium of the corpus callosum and posterior limb of the internal capsule. Moreover, a positive correlation between FA and anhedonia was found in the superior longitudinal fasciculus as well as a negative correlation in the cingulum. Then, the analysis of the anxiety and diffusion metric revealed that increased anxiety was associated with greater FA values in genu and splenium of corpus callosum, anterior corona radiata and posterior thalamic radiation. Finally, the motor retardation analysis showed a correlation between increased Widlöcher depressive retardation scale scores and reduced FA in the body and genu of the corpus callosum, fornix, and superior striatum. Through this twofold approach (categorical and phenotypic), this study has underlined the need to move forward to a symptom-based research area of biomarkers, which help to understand the pathophysiology of mood depressive disorders and to stratify precise phenotypes of depression with targeted therapeutic strategies [9]. This work was done with Centre Hospitalier Guillaume Régnier, Academic Psychiatry Department, 35703 Rennes, France.

7.2.6.2. Structural connectivity analysis in treatment-resistant depression

Participants: Julie Coloigner, Isabelle Corouge, Christian Barillot.

Depressive disorder is characterized by a profound dysregulation of affect and mood as well as additional abnormalities including cognitive dysfunction, insomnia, fatigue and emotional disturbance. Converging evidence shows that a dysfunction in prefrontal-subcortical circuits is associated with depressive state. However, the process of treatment resistance was poorly studied. One study of functional magnetic resonance imaging has reported more disrupted connectivity in prefrontal areas and in thalamus for resistant (R) group (Lui et al., 2011). These observations suggest a modification of functional connectivity in the prefrontal-subcortical circuits in the R patients. Using graph theory-based analysis, we examined white matter changes in the organization of networks in R patients compared with non-resistant (NR) group. We revealed 15 areas with significant density differences in R patients compared to NR subjects. The NR depression seems associated with decreased connectivity among distributed limbic areas, particularly in the ACC and in basal ganglia. However, the R patients exhibit a reduced connectivity in anterior limb of internal capsule and genu of corpus callosum compared with NR patients. Combined with previous studies, which described a widespread disruption in prefrontal-subcortical networks, this result suggests a more important connectivity decrease in the frontal cortex, as well as a smaller reduction in the limbic circuit for the patients with pejorative outcome. These results were consistent with connectivity studies, which suggested that the degree of disruption could influence the resistance severity and that two distinct networks could be implicated in NR end R depression. [27].

7.2.7. Prenatal exposure

7.2.7.1. Prenatal exposure to glycol ethers and motor inhibition function evaluated by functional MRI at the age of 10 to 12years in the PELAGIE mother-child cohort

Participants: Élise Bannier, Christian Barillot.

Pregnant women are ubiquitously exposed to organic solvents, such as glycol ethers. Several studies suggest potential developmental neurotoxicity following exposure to glycol ethers with a lack of clarity of possible brain mechanisms. We investigated the association between urinary levels of glycol ethers of women during early pregnancy and motor inhibition function of their 10- to 12-year-old children by behavioral assessment and brain MR imaging. Prenatal urinary levels of two glycol ether metabolites were associated with poorer Go/No-Go task performance. Differential activations were observed in the brain motor inhibition network in relation with successful inhibition, but not with cognitive demand. Nevertheless, there is no consistence between performance indicators and cerebral activity results. Other studies are highly necessary given the ubiquity of glycol ether exposure [5].

This work is done in collaboration with Fabienne Pelé and Cécile Chevrier (IRSET). Anne Claire Binter defended her PhD in December 2019 supervised by Fabienne Pelé, Cécile Chevrier and Élise Bannier. t

7.2.7.2. Effect of prenatal organic solvent exposure on structural connectivity at childhood

Participants: Julie Coloigner, Élise Bannier, Jean-Christophe Ferré, Christian Barillot.

Glycol ethers are part of organic solvents. They are used in industry and at home during manufacturing or usage of products such as paints, cleaning agents and cosmetics. The specific detection of subtle, low-dose effects of early-life exposure to these solvents on neuropsychological performance in children is a trendy subject of investigation. Neuroimaging allows looking into brain function and identifying different cerebral connections that may be affected by these neurotoxicants. In this paper, we investigated the specific effects of prenatal low-level exposure to different glycol ethers, on brain development of children between 10 and 12 years old. Based on previous studies suggesting cognitive disabilities in the attention, inhibition and working memory, we proposed a structural connectivity analysis using graph theory restricted to the regions involved in these functions. Our results suggest a possible relationship between the attention, working memory and inhibition and prenatal exposure to specific glycol ethers, such as ethoxyacetic acid, ethoxyethoxyacetic acid and 2-butoxyacetic acid [28].

7.2.8. Cognitive food-choice task

7.2.8.1. Implemenation of a new food picture database in the context of fMRI and visual cognitive food-choice task in healthy volunteers

Participant: Élise Bannier.

This pilot study aimed at implementing a new food picture database in the context of functional magnetic resonance imaging (fMRI) cognitive food-choice task, with an internal conflict or not, in healthy normal-weight adults. The fMRI analyses showed that the different liking foods (i.e. foods with different hedonic appraisals) condition elicited the activation of dorsal anterior cingulate cortex, involved in internal conflict monitoring, whereas similar liking (ie, foods with similar hedonic appraisals) condition did not, and that low-energy (LE) food choice involved high-level cognitive processes with higher activation of the hippocampus (HPC) and fusiform gyrus compared to high-energy (HE) food choice. Overall, this pilot study validated the use of the food picture database and fMRI-based procedure assessing decision-making processing during a food choice cognitive task with and without internal conflict[15].

This work was done in collaboration with Yentl Gautier, Paul Meurice, Yann Serrand, Nicolas Coquery Romain Moirand and David Val-Laillet from the NuMeCan Institute (Nutrition Metabolisms Cancer, UMR 1241, Inserm - Université de Rennes 1) and INRA.

7.3. Research axis 3: Management of Information in Neuroimaging

In the context of population imaging, we have made progress in three main areas this year. First we were involved in the development of infrastructures for open science with OpenAIRE, we also participated in the collaborative definition of standards that will ensure that infrastructures remain interoperable. Finally, we started a research new axis looking at how variations in analytical pipelines impact neuroimaging results (i.e. analytic variability).

7.3.1. Infrastructures

7.3.1.1. Open research: linking the bits and pieces with OpenAIRE-connect

Participants: Camille Maumet, Christian Barillot, Xavier Rolland.

Open research is growing in neuroimaging. The community — supported by funders who want best use of public funding but also by the general public who wants more transparent and participatory research practices — is constantly expanding online resources including: data, code, materials , tutorials, etc. This trend will likely amplify in the future and is also observed in other areas of experimental sciences. Open resources are typically deposited in dedicated repositories that are tailored to a particular type of artefact. While this is best practice, it makes it difficult to get the big picture: artefacts are scattered across the web in a multitude of databases. Although one could claim that the publication is here to link all related artefacts together, it is not machine-readable and does not me toallow searching for artefacts using filters (e.g. all datasets created in relation with a given funder). We presented OpenAIRE-connect, an overlay platform that links together research resources stored on the web: https://beta.ni.openaire.eu/ [45].

This work was done in collaboration with Dr. Sorina Caramasu-Pop and Axel Bonnet from Creatis in Lyon and with collaborators of the OpenAIRE-Connect project.

7.3.2. Standardisation and interoperability

7.3.2.1. The best of both worlds: using semantic web with JSON-LD. An example with NIDM-Results & Datalad

Participant: Camille Maumet.

The Neuroimaging data model (NIDM-Results) provides a harmonised representation for fMRI results reporting using Semantic Web technologies. While those technologies are particularly well suited for aggregation across complex datasets, using them can be costly in terms of initial development time to generate and read the corresponding serialisations. While the technology is machine accessible, it can be difficult to comprehend by humans. This hinders adoption by scientific communities and by software developers used to more-lightweight data-exchange formats, such as JSON. JSON-LD: a JSON representation for semantic graphs ("JSON-LD 1.1" n.d.) was created to address this limitation and recent extensions to the specification allow creating JSON-LD documents that are structured more similar to simple JSON. This representation is simultaneously readable by a large number of JSON-based applications and by Semantic Web tools. Here we review our work on building a JSON-LD representation for NIDM-Results data and exposing it to Datalad, a data-management tool suitable for neuroimaging datasets with built-in support for metadata extraction and search [44].

This work was done in collaboration with Prof. Michael Hanke from Institute of Neuroscience and Medicine in Julich and with members of the INCF.

7.3.2.2. Tools for FAIR Neuroimaging Experiment Metadata Annotation with NIDM Experiment

Participant: Camille Maumet.

Acceleration of scientific discovery relies on our ability to effectively use data acquired by consortiums and/or across multiple domains to generate robust and replicable findings. Efficient use of existing data relies on metadata being FAIR1 - Findable, Accessible, Interoperable and Reusable. Typically, data are shared using formats appropriate for the specific data types with little contextual information. Therefore, scientists looking to reuse data must contend with data originating from multiple sources, lacking complete acquisition information and often basic participant information (e.g. sex, age). What is required is a rich metadata standard that allows annotation of participant and data information throughout the experiment workflow, thereby allowing consumers easy discovery of suitable data. The Neuroimaging Data Model (NIDM)2 is an ongoing effort to represent, in a single core technology, the different components of a research activity, their relations, and derived data provenance3. NIDM-Experiment (NIDM-E) is focused on experiment design, source data descriptions, and information on the participants and acquisition information. In this work we report on annotation tools developed as part of the PyNIDM4 application programming interface (API) and their application to annotating and extending the BIDS5 versions of ADHD2006 and ABIDE7 datasets hosted in DataLad[40].

This work was led by Dr David Keator from UCI Irvine and done in collaboration with members of the INCF.

7.3.3. Quantifying analytic variability

7.3.3.1. Exploring the impact of analysis software on task fMRI results

Participant: Camille Maumet.

A wealth of analysis tools are available to fMRI researchers in order to extract patterns of task variation and, ultimately, understand cognitive function. However, this 'methodological plurality' comes with a drawback. While conceptually similar, two different analysis pipelines applied on the same dataset may not produce the same scientific results. Differences in methods, implementations across software packages, and even operating systems or software versions all contribute to this variability. Consequently, attention in the field has recently been directed to reproducibility and data sharing. Neuroimaging is currently experiencing a surge in initiatives to improve research practices and ensure that all conclusions inferred from an fMRI study are replicable. In this work, our goal is to understand how choice of software package impacts on analysis results. We use publically shared data from three published task fMRI neuroimaging studies, reanalyzing each study using the three main neuroimaging software packages, AFNI, FSL and SPM, using parametric and nonparametric inference. We obtain all information on how to process, analyze, and model each dataset from the publications. We make quantitative and qualitative comparisons between our replications to gauge the scale of variability in our results and assess the fundamental differences between each software package. While qualitatively we find broad similarities between packages, we also discover marked differences, such as Dice similarity coefficients ranging from 0.000-0.743 in comparisons of thresholded statistic maps between software. We discuss the challenges involved in trying to reanalyse the published studies, and highlight our own efforts to make this research reproducible [6].

This work was done in collaboration with Alexander Bowring and Prof. Thomas Nichols from the Oxford Big Data Institute in the UK.

FLUMINANCE Project-Team

6. New Results

6.1. Fluid motion estimation

6.1.1. Stochastic uncertainty models for motion estimation

Participants: Musaab Khalid Osman Mohammed, Etienne Mémin.

This work is concerned with the design of motion estimation technique for image-based river velocimetry. The method proposed is based on an advection diffusion equation associated to the transport of large-scale quantity with a model of the unresolved small-scale contributions. Additionally, since there is no ground truth data for such type of image sequences, a new evaluation method to assess the results has been developed. It is based on trajectory reconstruction of few Lagrangian particles of interest and a direct comparison against their manually-reconstructed trajectories. The new motion estimation technique outperformed traditional optical flow and PIV-based methods used in hydrology [23]. This study has been performed within the PhD thesis of Musaab Khalid and through a collaboration with the Irstea Lyon hydrology research group (HHLY).

6.1.2. Development of an image-based measurement method for large-scale characterization of indoor airflows

Participants: Dominique Heitz, Etienne Mémin, Romain Schuster.

The goal is to design a new image-based flow measurement method for large-scale industrial applications. From this point of view, providing in situ measurement technique requires: (i) the development of precise models relating the large-scale flow observations to the velocity; (ii) appropriate large-scale regularization strategies; and (iii) adapted seeding and lighting systems, like Hellium Filled Soap Bubles (HFSB) and led ramp lighting. This work conducted within the PhD of Romain Schuster in collaboration with the compagny ITGA has started in february 2016. The first step has been to evaluate the performances of a stochastic uncertainty motion estimator when using large scale scalar images, like those obtained when seeding a flow with smoke. The PIV characterization of flows on large fields of view requires an adaptation of the motion estimation method from image sequences. The backward shift of the camera coupled to a dense scalar seeding involves a large scale observation of the flow, thereby producing uncertainty about the observed phenomena. By introducing a stochastic term related to this uncertainty into the observation term, we obtained a significant improvement of the estimated velocity field accuracy. The technique was validated on a mixing layer in a wind tunnel for HFSB and smoke tracers [39] and applied on a laboratory fume-hood [26], [30], [43]. This study demonstrated the feasibility of conducting on-site large-scale image-based measurements for indoor airflows characterization. The technique was also assessed in an outdoor flow

6.1.3. 3D flows reconstruction from image data

Participants: Dominique Heitz, Etienne Mémin.

Our work focuses on the design of new tools for the estimation of 3D turbulent flow motion in the experimental setup of Tomo-PIV. This task includes both the study of physically-sound models on the observations and the fluid motion, and the design of low-complexity and accurate estimation algorithms. This year, we continued our investigation on the problem of efficient volume reconstruction via ensemble assimilation scheme. We have proposed a novel method for volumetric velocity reconstruction exploring the locality of 3D object space. Under this formulation the velocity of local patch was sought to match the projection of the particles within the local patch in image space to the image recorded by camera. The core algorithm to solve the matching problem is an instance-based estimation scheme that can overcome the difficulties of optimization originated from the nonlinear relationship between the imageintensity residual and the volumetric velocity. The proposed method labeled as Lagrangian Particle ImageVelocimetry (LaPIV) is quantitatively evaluated with synthetic particle image data. The promising results indicated the potential application of LaPIV to a large variety of volumetric velocity reconstruction problems [44].

6.2. Tracking, Data assimilation and model-data coupling

6.2.1. Optimal control techniques for the coupling of large scale dynamical systems and image data

Participants: Mohamed Yacine Ben Ali, Pranav Chandramouli, Dominique Heitz, Etienne Mémin, Gilles Tissot.

In this axis of work, we explore the use of optimal control techniques for the coupling of Large Eddies Simulation (LES) techniques and 2D image data. The objective is to reconstruct a 3D flow from a set of simultaneous time resolved 2D image sequences visualizing the flow on a set of 2D planes enlightened with laser sheets. This approach is experimented on shear layer flows and on wake flows generated on the wind tunnel of Irstea Rennes. Within this study we aim to explore techniques to enrich large-scale dynamical models by the introduction of uncertainty terms or through the definition of subgrid models from the image data. This research theme is related to the issue of turbulence characterization from image sequences. Instead of predefined turbulence models, we aim here at tuning from the data the value of coefficients involved in traditional LES subgrid models. A 4DVar assimilation technique based on the numerical code Incompact3D has been implemented for that purpose to control the inlet and initial conditions in order to reconstruct a turbulent wake flow behind an unknown obstacle [21]. We extended this first data assimilation technique to control the subgrid parameters. This study is performed in collaboration with Sylvain Laizet (Imperial College). In another axis of research, in collaboration with the CSTB Nantes centre and within the PhD of Yacine Ben Ali we will explore the definition of efficient data assimilation schemes for wind engineering. The goal is here to couple Reynolds average model to pressure data at the surface of buildings. The final purpose will consist in proposing improved data-driven simulation models for architects.

6.2.2. Ensemble variational data assimilation of large-scale dynamics with uncertainty

Participant: Etienne Mémin.

Estimating the parameters of geophysical dynamic models is an important task in Data Assimilation (DA) technique used for forecast initialization and reanalysis. In the past, most parameter estimation strategies were derived by state augmentation, yielding algorithms that are easy to implement but may exhibit convergence difficulties. The Expectation-Maximization (EM) algorithm is considered advantageous because it employs two iterative steps to estimate the model state and the model parameter separately. In this work, we propose a novel ensemble formulation of the Maximization step in EM that allows a direct optimal estimation of physical parameters using iterative methods for linear systems. This departs from current EM formulations that are only capable of dealing with additive model error structures. This contribution shows how the EM technique can be used for dynamics identification problem with a model error parameterized as arbitrary complex form. The proposed technique is used for the identification of stochastic subgrid terms that account for processes unresolved by a geophysical fluid model. This method, along with the augmented state technique, has been evaluated to estimate such subgrid terms through high resolution data. Compared to the augmented state technique, our method is shown to yield considerably more accurate parameters. In addition, in terms of prediction capacity, it leads to smaller generalization error as caused by the overfitting of the trained model on presented data and eventually better forecasts [29].

6.2.3. Reduced-order models for flows representation from image data

Participants: Dominique Heitz, Etienne Mémin, Gilles Tissot.

During the PhD thesis of Valentin Resseguier we have proposed a new decomposition of the fluid velocity in terms of a large-scale continuous component with respect to time and a small-scale non continuous random component. Within this general framework, an uncertainty based representation of the Reynolds transport theorem and Navier-Stokes equations can be derived, based on physical conservation laws. This physically relevant stochastic model has been applied in the context of POD-Galerkin methods. This uncertainty modeling methodology provides a theoretically grounded technique to define an appropriate subgrid tensor as well as drift correction terms. The pertinence of this stochastic reduced order model has been successfully assessed

on several wake flows at different Reynold number. It has been shown to be much more stable than the usual reduced order model construction techniques. Beyond the definition of a stable reduced order model, the modeling under location uncertainty paradigm offers a unique way to analyse from the data of a turbulent flow the action of the small-scale velocity components on the large-scale flow. Regions of prominent turbulent kinetic energy, direction of preferential diffusion, as well as the small-scale induced drift can be identified and analyzed to decipher key players involved in the flow. This study has been published in the Journal of Fluid Mechanics [15]. Note that these reduced order models can be extended to a full system of stochastic differential equations driving all the temporal modes of the reduced system (and not only the small-scale modes). This full stochastic system has been evaluated on wake flow at moderate Reynolds number. For this flow the system has shown to provide very good uncertainty quantification properties as well as meaningful physical behavior with respect to the simulation of the neutral modes of the dynamics. This study is pursued within a strong collaboration with the industrial partner: SCALIAN

6.2.4. Learning of the dynamics of large scale geophysical systems using semi-group theory for data assimilation

Participants: Etienne Mémin, Gilles Tissot.

The goal of this study is to propose new ensemble data assimilation methodologies to estimate oceanic and turbulent flows. In classical methods, from a distribution of initial conditions, an ensemble of simulations are computed and used for estimation. Ideally, from this solution, a new ensemble has to be generated to refine the estimation. However, due to large numerical costs and operational constraints, this iterative procedure is in practice intractable. In order to improve actual performances, we propose to take these limitations into account and to develop new methodologies able to better take advantage of the information contained in the ensemble and in the dynamical model. More precisely, we propose to learn the non-linear dynamical features of the system and to be able to reproduce it without having to run a new simulation. The formalism is based on two concepts: i) the reproducing kernel Hilbert spaces (RKHS) that are a basis of smooth functions in the phase space giving interpolatory properties ii) the Koopman operator, that is an infinite-dimensional operator able to propagate in time any observable of the phase space. These two elements allow to define a rigorous framework in which hypothesis classically done in ensemble methods appear naturally. Thus, classical methods enter in a special case of this new formalism, that allows us to generalise them in a way to improve the learning of the non-linear dynamical system. Numerical tests are performed using the Ginzburg-Landau equation and a quasi-geostrophic flow model.

6.2.5. Estimation and control of amplifier flows

Participant: Gilles Tissot.

Estimation and control of fluid systems is an extremely hard problem. The use of models in combination with data is central to take advantage of all information we have on the system. Unfortunately all flows do not present the same physical and mathematical behaviour, thus using models and methodologies specialised to the flow physics is necessary to reach high performances.

A class of flows, denoted "oscillator flows", are characterised by unstable modes of the linearised operator. A consequence is the dominance of relatively regular oscillations associated with a nonlinear saturation. Despite the non-linear behaviour, associated structures and dynamical evolution are relatively easy to predict. Canonical configurations are the cylinder wake flow or the flow over an open cavity.

By opposition to that, "amplifier flows" are linearly stable with regard to the linearised operator. However, due to their convective nature, a wide range of perturbations are amplified in time and convected away such that it vanishes at long time. The consequence is the high sensitivity to perturbations and the broad band response that forbid a low rank representation. Jets and mixing layers show this behaviour and a wide range of industrial applications are affected by these broad band perturbations. It constitutes then a class of problems that are worth to treat separately since it is one of the scientific locks that render hard the transfer of methodologies existing in flow control and estimation to industrial applications.

There exists a type of models, that we will denote as "parabolised", that are able to efficiently represent amplifier flows. These models, such as parabolised stability equations and one-way Navier-Stokes propagate, in the frequency domain, hydrodynamic instability waves over a given turbulent mean flow. We can note that these models, by their structure, give access to a natural experimental implementation. They are an ingredient adapted to represent the system, but have a mathematical structure strongly different from the dynamical models classically used in control and data assimilation. It is then important to develop new methodologies of control, estimation and data assimilation with these models to reach our objectives. Moreover, inventing new models by introducing the modelling under location uncertainties in these parabolised models will be perfectly adapted to represent the evolution and the variability of an instability propagating within a turbulent flow. It will be consistent with actual postprocessing of experimental data performed in similar flow configurations.

6.3. Analysis and modeling of turbulent flows and geophysical flows

6.3.1. Geophysical flows modeling under location uncertainty

Participants: Werner Bauer, Pranav Chandramouli, Long Li, Etienne Mémin.

In this research axis we have devised a principle to derive representation of flow dynamics under location uncertainty. Such an uncertainty is formalized through the introduction of a random term that enables taking into account large-scale approximations or truncation effects performed within the dynamics analytical constitution steps. Rigorously derived from a stochastic version of the Reynolds transport theorem [9], this framework, referred to as modeling under location uncertainty (LU), encompasses several meaningful mechanisms for turbulence modeling. It indeed introduces without any supplementary assumption the following pertinent mechanisms for turbulence modeling: (i) a dissipative operator related to the mixing effect of the large-scale components by the small-scale velocity; (ii) a multiplicative noise representing small-scale energy backscattering; and (iii) a modified advection term related to the so-called *turbophoresis* phenomena, attached to the migration of inertial particles in regions of lower turbulent diffusivity.

In a series of papers we have shown how LU modeling can be applied to provide stochastic representations of a variety of classical geophysical flows dynamics [12], [13], [14]. Numerical simulations and uncertainty quantification have been performed on Quasi Geostophic approximation (QG) of oceanic models. It has been shown that LU leads to remarkable estimation of the unresolved errors opposite to classical eddy viscosity based models. The noise brings also an additional degree of freedom in the modeling step and pertinent diagnostic relations and variations of the model can be obtained with different scaling assumptions of the turbulent kinetic energy (i.e. of the noise amplitude). For a wind forced QG model in a square box, which is an idealized model of north-Atlantic circulation, we have shown that for different versions of the noise the QG LU model leads to improve long-terms statistics when compared to classical large-eddies simulation strategies. For a QG model we have demonstrated that the LU model allows conserving the global energy. We have also shown numerically that Rosby waves were conserved and that inhomogeneity of the random component triggers secondary circulations. This feature enabled us to draw a formal bridge between a classical system describing the interactions between the mean current and the surface waves and the LU model in which the turbophoresis advection term plays the role of the classical Stokes drift.

Supported by funding from Inria-Mitacs Globalink, we hosted Ruediger Brecht, PhD student at Memorial University of Newfoundland, Canada, for a period of 3 months (May to August) in the Fluminance group. During his stay, Ruediger Brecht worked on the incorporation of a stochastic representation of the small-scale velocity component of a fluid flow in a variational integrator for the rotating shallow-water equations on the sphere, already developed within the first part of its PhD work. This work was based on an ongoing study in the group on a stochastic Quasi-geostrophic model and followed a series of works performed in the Fluminance group to define stochastic geophysical flow dynamics.

6.3.2. Large eddies simulation models under location uncertainty

Participants: Mohamed Yacine Ben Ali, Pranav Chandramouli, Dominique Heitz, Etienne Mémin, Gilles Tissot.

The models under location uncertainty recently introduced by Mémin (2014) [9] provide a new outlook on LES modeling for turbulence studies. These models are derived from a stochastic transport principle. The associated stochastic conservation equations are similar to the filtered Navier- Stokes equation wherein we observe a sub-grid scale dissipation term. However, in the stochastic version, an extra term appears, termed as "velocity bias", which can be treated as a biasing/modification of the large-scale advection by the small scales. This velocity bias, introduced artificially in the literature, appears here automatically through a decorrelation assumption of the small scales at the resolved scale. All sub-grid contributions for the stochastic models are defined by the small-scale velocity auto-correlation tensor. This large scale modeling has been assed and compared to several classical large-scale models on a flow over a circular cylinder at Re 3900 and wall-bounded flows. For all these flows the modeling under uncertainty has provided better results than classical large eddies simulation models. Within the PhD of Yacine Ben Ali we will explore with the CSTB Nantes centre the application of such models for the definition of Reynolds average simulation (RANS) models for wind engineering applications.

6.3.3. Variational principles for structure-preserving discretizations in stochastic fluid dynamics

Participants: Werner Bauer, Long Li, Etienne Mémin.

The overarching goal of this interdisciplinary project is to use variational principles to derive deterministic and stochastic models and corresponding accurate and efficient structure preserving discretizations and to use these schemes to obtain a deeper understanding of the conservation laws of the stochastic fluid dynamics investigated. The newly developed systematic discretization framework is based on discrete variational principles whose highly structured procedures shall be exploited to develop a general software framework that applies automatic code generation. This project will first provide new stochastic fluid models and suitable approximations, with potential future applications in climate science using the developed methods to perform accurate long term simulations while quantifying the solutions uncertainties. The generality of our approach addresses also other research areas such as electrodynamics (EDyn), magnetohydrodynamics (MHD), and plasma physics.

6.3.4. Stochastic compressible fluid dynamics

Participants: Etienne Mémin, Gilles Tissot.

Some work has been performed to extend the stochastic formulation under location uncertainty to compressible flows. The interest is to extend the formulation on the one hand to compressible fluids (for instability mechanisms involved in areoacoustics for instance, or for thermal effects in mixing layers) and on the other hand to geophysical flows where the Boussinesq equation is not valid anymore (density variations due to temperature or salinity gradients). A theoretical study has been performed that opens the door to numerical validations. In particular a baroclinic torque term has been identified that could have major effects in some situations.

6.3.5. Stochastic hydrodynamic stability under location uncertainty

Participants: Etienne Mémin, Gilles Tissot.

In order to predict instability waves propagating within turbulent flows, eigenmodes of the linearised operator is not well suited since it neglects the effect of turbulent fluctuations on the wave dynamics. To cope this difficulty, resolvent analysis has become popular since it represents the response of the linearised operator to any forcing representing the generalised stress tensors. The absence of information on the non-linearity is a strong limitation of the method. In order to refine these models, we propose to consider a stochastic model under location uncertainty expressed in the Fourier domain, to linearise it around the corrected mean-flow and to study resulting eigenmodes. The stochastic part represents the effect of the turbulent field onto the instability wave. It allows to specify a structure of the noise and then to improve existing models. Improvements compared to the resolvent analysis have been found for turbulent channel flow data at $\mathfrak{R}_\tau = 180$. This work is in collaboration with André Cavalieri (Instituto Tecnologico de Aeronautica, SP, Brésil).

6.3.6. Singular and regular solutions to the Navier-Stokes equations (NSE) and relative turbulent models

Participants: Roger Lewandowski, Etienne Mémin, Benoit Pinier.

The common thread of this work is the problem set by J. Leray in 1934 : does a regular solution of the Navier-Stokes equations (NSE) with a smooth initial data develop a singularity in finite time, what is the precise structure of a global weak solution to the Navier-Stokes equations, and are we able to prove any uniqueness result of such a solution. This is a very hard problem for which there is for the moment no answer. Nevertheless, this question leads us to reconsider the theory of Leray for the study of the Navier-Stokes equations in the whole space with an additional eddy viscosity term that models the Reynolds stress in the context of large-scale flow modelling. It appears that Leray's theory cannot be generalized turnkey for this problem, so that things must be reconsidered from the beginning. This problem is approached by a regularization process using mollifiers, and particular attention must be paid to the eddy viscosity term. For this regularized problem and when the eddy viscosity has enough regularity, we have been able to prove the existence of a global unique solution that is of class C? in time and space and that satisfies the energy balance. Moreover, when the eddy viscosity is of compact support in space, uniformly in time, we recently shown that this solution converges to a turbulent solution to the corresponding Navier-Stokes equations,carried when the regularizing parameter goes to 0. These results are described in a paper published in JMAA [24]

In the framework of the collaboration with the University of Pisa (Italy), namely with Luigi Berselli collaboration, we considered the three dimensional incompressible Navier-Stokes equations with non stationary source terms chosen in a suitable space. We proved the existence of Leray-Hopf weak solutions and that it is possible to characterize (up to sub-sequences) their long-time averages, which satisfy the Reynolds averaged equations, involving a Reynolds stress. Moreover, we showed that the turbulent dissipation is bounded by the sum of the Reynolds stress work and of the external turbulent fluxes, without any additional assumption, than that of dealing with Leray-Hopf weak solutions. This is a very nice generalisation to non stationnary source terms of a famous results by Foais. IN the same work, we also considered ensemble averages of solutions, associated with a set of different forces and we proved that the fluctuations continue to have a dissipative effect on the mean flow. These results have been published in Nonlinearity [19]. These results have been extended in the framework of POD for reduced models in [18].

In [55] we have shown the existence of a solution to a 1D Reynolds Averaged Navier-Stokes vertical model suitable in the atmospheric boundary layer, under suyitable assumption on the data. The paper is received for publication in thje journal Pure and Applied Functional Analysis (PAFA).

We also have introduced a turbulence model including a backscatter term, which has the same structure as the Voigt model. The additional term is derived in certain specific regimes of the flow, such as the convergence to stable statistical states. We get estimates for the velocity v in $L^\infty_t H^1_x \cap W^{1,2}_t H^{1/2}_x$, that allow us to prove the existence and uniqueness of a regular-weak solutions (v, p) to the resulting system, for a given fixed eddy viscosity. We then prove a structural compactness result that highlights the robustness of the model. This allows us to pass to the limit in the quadratic source term in the equation for the turbulent kinetic energy k, which yields the existence of a weak solution to the corresponding Reynolds Averaged Navier-Stokes system satisfied by (v, p, k). These results are writen in [47], a paper which is under revision in Non Linear Analysis.

Another study in collaboration with B. Pinier, P. Chandramouli and E. Memin has been undertaken. This work takes place within the context of the PhD work of B. Pinier. We have tested the performances of an incompressible turbulence Reynolds-Averaged Navier-Stokes one-closure equation model in a boundary layer, which requires the determination of the mixing length l. A series of direct numerical simulation have been performed, with flat and non trivial topographies, to obtain by interpolation a generic formula l = l(Re6, z), Re6 being the frictional Reynolds number, and z the distance to the wall. Numerical simulations have been carried out at high Reynolds numbers with this turbulence model, in order to discuss its ability to properly reproduce the standard profiles observed in neutral boundary layers, and to assess its advantages, its disadvantages and its limits. We also proceeded to a mathematical analysis of the model.

6.3.7. Stochastic flow model to predict the mean velocity in wall bounded flows

Participants: Roger Lewandowski, Etienne Mémin, Benoit Pinier.

To date no satisfying model exists to explain the mean velocity profile within the whole turbulent layer of canonical wall bounded flows. We propose a modification of the velocity profile expression that ensues from the stochastic representation of fluid flows dynamics proposed recently in the group and referred to as "modeling under location uncertainty". This framework introduces in a rigorous way a subgrid term generalizing the eddy-viscosity assumption and an eddy-induced advection term resulting from turbulence inhomogeneity. This latter term gives rise to a theoretically well-grounded model for the transitional zone between the viscous sublayer and the turbulent sublayer. An expression of the small-scale velocity component is also provided in the viscous zone. Numerical assessment of the results have been performed for turbulent boundary layer flows, pipe flows and channel flows at various Reynolds numbers [25][17].

6.3.8. Numerical and experimental image and flow database

Participants: Pranav Chandramouli, Dominique Heitz.

The goal was to design a database for the evaluation of the different techniques developed in the Fluminance group. The first challenge was to enlarge a database mainly based on two-dimensional flows, with three-dimensional turbulent flows. Synthetic image sequences based on homogeneous isotropic turbulence and on circular cylinder wake have been provided. These images have been completed with time resolved Particle Image Velocimetry measurements in wake and mixing layers flows. This database provides different realistic conditions to analyse the performance of the methods: time steps between images, level of noise, Reynolds number, large-scale images. The second challenge was to carry out orthogonal dual plane time resolved stereoscopic PIV measurements in turbulent flows. The diagnostic employed two orthogonal and synchronized stereoscopic PIV measurements to provide the three velocity components in planes perpendicular and parallel to the streamwise flow direction. These temporally resolved planar slices observations have been be used within a 4DVar assimilation technique, to reconstruct three-dimensional turbulent flows from data. The third challenge was to carry out a time resolved tomoPIV experiments in a turbulent wake flow. This work has been submitted to the Journal of Computational Physics.

6.3.9. Fast 3D flow reconstruction from 2D cross-plane observations

Participants: Pranav Chandramouli, Dominique Heitz, Etienne Mémin.

We proposed a computationally efficient flow reconstruction technique, exploiting homogeneity in a given direction, to recreate three dimensional instantaneous turbulent velocity fields from snapshots of two dimension planar fields. This methodology, termed as "snapshot optimisation" or SO, enables to provide 3D data-sets for studies which are currently restricted by the limitations of experimental measurement techniques. The SO method aims at optimising the error between an inlet plane with a homogeneous direction and snap-shots, obtained over a sufficient period of time, on the observation plane. The observations are carried out on a plane perpendicular to the inlet plane with a shared edge normal to the homogeneity direction. The method is applicable to all flows which display a direction of homogeneity such as cylinder wake flows, channel flow, mixing layer, and jet (axi-symmetric). The ability of the method is assessed with two synthetic data-sets, and three experimental PIV data-sets. A good reconstruction of the large-scale structures are observed for all cases. This study has been published in the journal "Experiments in Fluids" [21].

6.4. Visual servoing approach for fluid flow control

6.4.1. A state space representation for the closed-loop control of shear flows

Participants: Johan Carlier, Christophe Collewet.

The goal of this study is to develop a generic state representation for the closed-loop control of shear flows. We assume that the actuator acts at the boundaries. Our approach is based on a linearization of the Navier-Stokes equations around the desired state. Particular care was paid to the discrete approximation of the linear model to design a well-conditioned and accurate state matrix describing time evolution of disturbances evolving in parallel shear flow as long as these disturbances remain sufficiently small. A state matrix representation is obtained for the periodic channel flow and the spatially developing mixing layer flow. This approach has been validated through the representativity of our model in terms of linear stability. This work has been presented to the French Mechanics Congress CFM'2019 (https://hal.inria.fr/hal-02283161) [36].

6.4.2. Closed-loop control of a spatially developing shear layer

Participants: Christophe Collewet, Johan Carlier.

This study aims at controling one of the prototypical flow configurations encountered in fluid mechanics: the spatially developing turbulent shear layer occuring between two parallel incident streams with different velocities. Our goal is to maintain the shear-layer in a desired state and thus to reject upstream perturbations. In our conference IFAC paper (https://hal.inria.fr/hal-01514361) we focused on perturbations belonging to the same space that the actuators, concretely that means that we were only able to face perturbations of the actuator itself, like failures of the actuator. This year we enlarged this result to purely exogenous perturbations, in term of magnitude as well as in term of spatial dispersion. An optimal control law has been derived to minimize the influence of the perturbation on the flow. To do that, an on-line estimation of the perturbation (magnitude and spatial dispersion) has been developed to lead to an adaptive control law. Simple conditions to ensure the local asymptotic stability of the whole scheme have been derived. This work has been also presented to the French Mechanics Congress CFM'2019 (https://hal.archives-ouvertes.fr/hal-02189111) [37].

6.4.3. Design of a DBD plasma actuator for closed-loop control

Participants: Johan Carlier, Christophe Collewet.

The goal of this study is to design a DBD plasma actuator for closed-loop control. This kind of actuator is widely used in the flow control community however, it is more appropriate to force a flow than to control it. Indeed, to control a flow under a closed-loop fashion, the action must be proportional to the control signal provided by the control law. It is unfortunately not the case with these actuators. We have modified the classical DBD plasma actuator so that the action is almost a linear fonction of the control signal. Our approach have been validated by a prototype and by first experiments.

6.5. Coupled models in hydrogeology

6.5.1. Reactive transport in multiphase flow

Participant: Jocelyne Erhel.

Groundwater resources are essential for life and society, and should be preserved from contamination. Pollutants are transported through the porous medium and a plume can propagate. Reactive transport models aims at simulating this dynamic contamination by coupling advection dispersion equations with chemistry equations. If chemistry is at thermodynamic equilibrium, then the system is a set of partial differential and algebraic equations (PDAE). Space discretization leads to a semi-discrete DAE system which should be discretized in time. An explicit time scheme allows an easy decoupling of transport and chemistry, but very small timesteps should be taken, leading to a very large CPU time. Therefore, an implicit time scheme is preferred, coupling transport and chemistry in a nonlinear system. The special structure of linearized systems can be used in preconditioned Newton-Krlyov methods in order to improve efficiency. Some experiments illustrate the methodology and show also the need for an adaptive timestep and a control of convergence in Newton's iterations.

This work was presented at a workshop [31].

6.5.2. Characterizations of Solutions in Geochemistry at equilibrium

Participant: Jocelyne Erhel.

Geochemistry at thermodynamic equilibrium involves aqueous reactions and mineral precipitation or dissolution. Quantities of solute species are assumed to be strictly positive, whereas those of minerals can vanish. The mathematical model is expressed as the minimization of Gibbs energy subject to positivity of mineral quantities and conservation of mass. Optimality conditions lead to a complementarity problem. We show that, in the case of a dilute solution, this problem can also be considered as optimality conditions of another minimization problem, subject to inequality constraints. This new problem is easier to handle, both from a theoretical and a practical point of view. Then we define a partition of the total quantities in the mass conservation equation. This partition builds a precipitation diagram such that a mineral is either precipitated or dissolved in each subset. We propose a symbolic algorithm to compute this diagram. Simple numerical examples illustrate our methodology.

This work was published in the journal Computational Geosciences [22] and presented at an international conference [38].

6.5.3. Mathematical models of kinetic reactions in geochemistry

Participants: Jocelyne Erhel, Bastien Hamlat.

In geochemistry, kinetic reactions can lead to the appearance or disappearance of minerals or gas. We defined two mathematical models based first on a differential inclusion system and second on a projected dynamical system. We proposed a regularization process for the first model and a projection algorithm for the second one.

This work, supported by IFPEN, was presented at a conference [39] and a workshop [32].

6.6. Sparse Linear solvers

6.6.1. Parallel GMRES

Participant: Jocelyne Erhel.

Sparse linear systems $Ax = b$ arise very often in computational science and engineering. Krylov methods are very efficient iterative methods, and restarted GMRES is a reference algorithm for non-symmetric systems. A first issue is to ensure a fast convergence, by preconditioning the system with a matrix M. Preconditioning must reduce the number of iterations, and be easy to solve. A second issue is to achieve high performance computing. The most time-consuming part in GMRES is to build an orthonormal basis V. With the Arnoldi process, many scalar products involve global communications. In order to avoid them, s-step methods have been designed to find a tradeoff between parallel performance and stability. Also, solving a system with the matrix M and for multiplying a vector by the matrix A should be efficient. A domain decomposition approach involves mainly local communications and is frequently used. A coarse grid correction, based on deflation for example, improves convergence. These techniques can be combined to provide fast convergence and fast parallel algorithms. Numerical results illustrate various issues and achievements.

This work was presented at an international conference (invited talk) [31].

GENSCALE Project-Team

7. New Results

7.1. Algorithms & Methods

7.1.1. SV genotyping

Participants: Dominique Lavenier, Lolita Lecompte, Claire Lemaitre, Pierre Peterlongo.

Structural variations (SV) are genomic variants of at least 50 base pairs (bp) that can be rearranged within the genome and thus can have a major impact on biological processes. Sequencing data from third generation technologies have made it possible to better characterize SVs. Although many SV callers have been published recently, there is no published method to date dedicated to genotyping SVs with this type of data. Variant genotyping consists in estimating the presence and ploidy or absence of a set of known variants in a newly sequenced individual. Thus, in this paper, we present a new method and its implementation, SVJedi, to genotype SVs with long reads. From a set of known SVs and a reference genome, our approach first generates local sequences representing the two possible alleles for each SV. Long read data are then aligned to these generated sequences and a careful analysis of the alignments consists in identifying only the informative ones to estimate the genotype for each SV. SVJedi achieves high accuracy on simulated and real human data and we demonstrate its substancial benefits with respect to other existing approaches, namely SV discovery with long reads and SV genotyping with short reads [23], [24], [35]. SVJedi is implemented in Python and available at https://github.com/llecompte/SVJedi.

7.1.2. Genome assembly of targeted organisms in metagenomic data

Participants: Wesley Delage, Fabrice Legeai, Claire Lemaitre.

In this work, we propose a two-step targeted assembly method tailored for metagenomic data, called MinYS (for MineYourSymbiont). First, a subset of the reads belonging to the species of interest are recruited by mapping and assembled *de novo* into backbone contigs using a classical assembler. Then an all-versus-all contig gap-filling is performed using a novel version of MindTheGap with the whole metagenomic dataset. The originality and success of the approach lie in this second step, that enables to assemble the missing regions between the backbone contigs, which may be regions absent or too divergent from the reference genome. The result of the method is a genome assembly graph in gfa format, accounting for the potential structural variations identified within the sample. We showed that MinYS is able to assemble the *Buchnera aphidicola* genome in a single contig in pea aphid metagenomic samples, even when using a divergent reference genome, it runs at least 10 times faster than classical *de novo* metagenomics assemblers and it is able to recover large structural variations co-existing in a sample. MinYS is a Python3 pipeline, distributed on github (https://github.com/cguyomar/MinYS) and as a conda package in the bio-conda repository [22].

7.1.3. SimkaMin: subsampling the kmer space for efficient comparative metagenomics

Participants: Claire Lemaitre, Pierre Peterlongo.

SimkaMin [12] is a quick comparative metagenomics tool with low disk and memory footprints, thanks to an efficient data subsampling scheme used to estimate Bray-Curtis and Jaccard dissimilarities. One billion metagenomic reads can be analyzed in less than 3 minutes, with tiny memory (1.09 GB) and disk ($\sim$0.3 GB) requirements and without altering the quality of the downstream comparative analyses, making of SimkaMin a tool perfectly tailored for very large-scale metagenomic projects.

7.1.4. Haplotype reconstruction: phasing co-localized variants

Participants: Mohammed Amin Madoui, Pierre Peterlongo.

In collaboration with Amin Madoui from the Genoscope (CEA), we develop a new methodology to reconstruct haplotypes or strain genomes directly from raw sequencing set of (metagenomic) reads. The goal is to propose long assembled sequences (i.e. complete genomes are not mandatory) such that each assembled sequence belongs to only one sequenced chromosome and is not a consensus of several similar sequences. Downstream, this enables to perform population genomics analyses.

The key idea is to use the DiscoSnp [10] output, detecting set of variant alleles that are co-localized on input reads or pairs of input reads. Then we finally reconstruct set of sequences that are as parsimonious as possible with those observations.

7.1.5. Finding all maximal perfect haplotype blocks in linear time

Participant: Pierre Peterlongo.

Recent large-scale community sequencing efforts allow at an unprecedented level of detail the identification of genomic regions that show signatures of natural selection. However, traditional methods for identifying such regions from individuals' haplotype data require excessive computing times and therefore are not applicable to current datasets. In 2019, Cunha et al. (Proceedings of BSB 2019) suggested the maximal perfect haplotype block as a very simple combinatorial pattern, forming the basis of a new method to perform rapid genome-wide selection scans. The algorithm they presented for identifying these blocks, however, had a worst-case running time quadratic in the genome length. It was posed as an open problem whether an optimal, linear-time algorithm exists. We gave two algorithms that achieve this time bound, one which is conceptually very simple and uses suffix trees and a second one using the positional Burrows-Wheeler Transform, that is very efficient also in practice [20].

7.1.6. Short read correction

Participant: Pierre Peterlongo.

We propose a new approach for the correction of NGS reads. This approach is based on the construction of a clean de Bruijn graph in which the correction is made at the contig level. In a second step, original reads are mapped on this graph, allowing to correct the original reads [16].

7.1.7. Large-scale kmer indexation

Participants: Téo Lemane, Pierre Peterlongo.

In the SeqDigger ANR project framework (see dedicated Section), we aim to index TB or PB of genomic sequences, assembled or not. The central idea is to assign any kmer (word of length k) to the set of indexed dataset it belongs to. For doing this we have proposed a method that improves one of the state of the art algorithm (HowDeSBT [38]) by optimizing the way kmers are counted and represented [36].

7.1.8. Proteogenomics workflow for the expert annotation of eukaryotic genomes

Participant: Pierre Peterlongo.

Accurate structural annotation of genomes is still a challenge, despite the progress made over the past decade. The prediction of gene structure remains difficult, especially for eukaryotic species, and is often erroneous and incomplete. In [15], we proposed a proteogenomics strategy, taking advantage of the combination of proteomics datasets and bioinformatics tools, to identify novel protein coding-genes and splice isoforms, to assign correct start sites, and to validate predicted exons and genes.

7.1.9. Gap-filling with linked-reads data

Participants: Anne Guichard, Fabrice Legeai, Claire Lemaitre, Arthur Le Bars, Pierre Peterlongo.

We develop a novel approach for filling assembly gaps with linked reads data (typically 10X Genomics technology). The approach is based on local assembly using our tool MindTheGap [9], and takes advantage of barcode information to reduce the input read set in order to reduce the de Bruijn graph complexity. The approach is applied to recover the genomic structure of a 1.3 Mb locus of interest in a dozen of re-sequenced butterfly genomes (*H. numata*) in the Supergene ANR project context.

7.1.10. Statistically Significant Discriminative Patterns Searching

Participants: Gwendal Virlet, Dominique Lavenier.

We propose a novel algorithm, called SSDPS, to discover patterns in two-class datasets. The algorithm, developed in collaboration with the LACODAM Inria team, owes its efficiency to an original enumeration strategy of the patterns, which allows to exploit some degrees of anti-monotonicity on the measures of discriminance and statistical significance. Experimental results demonstrate that the performance of the algorithm is better than others. In addition, the number of generated patterns is much less than the number of the other algorithms. An experiment on real data also shows that SSDPS efficiently detects multiple SNPs combinations in genetic data [27].

7.2. Optimisation

7.2.1. Chloroplast genome assembly

Participants: Sébastien Francois, Roumen Andonov, Dominique Lavenier.

This research focuses on the last two stages of *de novo* genome assembly, namely, scaffolding and gap-filling, and shows that they can be solved as part of a single optimization problem. Our approach is based on modeling genome assembly as a problem of finding a simple path in a specific graph that satisfies as many distance constraints as possible encoding the read-pair insert-size information. We formulate it as a mixed-integer linear programming (MILP) problem and apply an optimization solver to find the exact solutions on a benchmark of chloroplast genomes. We show that the presence of repetitions in the set of unitigs is the main reason for the existence of multiple equivalent solutions that are associated to alternative subpaths. We also describe two sufficient conditions and we design efficient algorithms for identifying these subpaths. Comparisons of the results achieved by our tool with the ones obtained with recent assemblers are also presented [11].

7.2.2. Integer Linear Programming for Metabolic Networks

Participants: Kerian Thuillier, Roumen Andonov.

Metabolic networks are a helpful tool to represent and study cell metabolisms. They contain information about every reaction occurring inside an organism. However, metabolic networks of poorly studied species are often incomplete. It is possible to complete these networks with knowledge of other well-known species.

In this study, we present a new linear programming approach for the problem of topological activation in metabolic networks based on flows and the Miller, Tucker and Zemlin (MTZ) formulation for solving the longest path problem. We developed a tool called *Flutampl* with AMPL (A Mathematical Programming Language). It returns optimal solutions for the hybrid completion directly from *sbml* files (the data format used for modelling metabolic networks) [37].

7.2.3. Integer Linear Programming for De novo Long Reads Assembly

Participants: Victor Epain, Roumen Andonov, Dominique Lavenier.

To tackle the de novo long read assembly problem, we investigate a new 2-step method based on integer linear programming. The first step orders the long reads and the second one generates a consensus sequence. Each step is based on a different IPL specification. In 2019, we focused on step 1: long reads are first compared to build an overlapping graph. Then we use integer linear programming to find the heaviest path in a graph $G = (V, E, \lambda)$, where V is the vertices set corresponding to the long reads, E the edge set associated to the overlaps between long reads and λ the overlap length. For large graph, V is partitioned into several parts, each one is solved independently, and the solutions are merged together. Preliminary experimentation show that bacteria assemblies can be successfully solved in a few minutes [31].

7.3. Experiments with the MinION Nanopore sequencer

7.3.1. Storing information on DNA Molecules

Participants: Dominique Lavenier, Emeline Roux, Ariane Badoual.

In 2019, we started a new research activity aiming to explore the possibility to use the DNA molecules as a storage medium. We designed a complete DNA storage system based on the Oxford Nanopore sequencing technology and performed several experimentations by sequencing several synthetized DNA fragments ranging from 500 to 1,000 bp. These sequences have been designed with ad-hoc coding to prevent specific sequencing errors of the nanopore technology such as indel errors in homo-polymer sequences [29] [34]. These real experiments demonstrate that a text encoded into the DNA alphabet, then synthetized into DNA molecules, sequenced with the MinION, and finally processed using bioinformatics approaches can be fully recovered [28].

7.3.2. Identification of bacterial strains

Participants: Jacques Nicolas, Emeline Roux, Grégoire Siekaniec, Clara Delahaye.

Our aim is to provide rapid algorithms for the identification of bacteria at the finest taxonomic level. We have developed an expertise in the use of the MinIon long read technology and have produced and assembled many genomes for lactic bacteria in cooperation with INRA STLO, which have been made available on the Microscope platform at Genoscope (http://www.genoscope.cns.fr/agc/microscope/). We have developed a first classifier that demonstrates the possibility to identify isolated strains with spaced seed indexing of the noisy long reads produced by the MinIon.

7.4. Benchmarks and Reviews

7.4.1. Evaluation of error correction tools for long Reads

Participants: Lolita Lecompte, Pierre Peterlongo.

Long read technologies, such as Pacific Biosciences and Oxford Nanopore, have high error rates (from 9% to 30%). Hence, numerous error correction methods have been recently proposed, each based on different approaches and, thus, providing different results. As this is important to assess the correction stage for downstream analyses, we designed the ELECTOR software, providing evaluation of long read correction methods. This software generates additional quality metrics compared to previous existing tools. It also scales to very long reads and large datasets and is compatible with a wide range of state-of-the-art error correction tools [17]. ELECTOR is freely available at https://github.com/kamimrcht/ELECTOR.

7.4.2. Evaluation of insertion variant callers on real human data

Participants: Wesley Delage, Claire Lemaitre.

Insertion variants are one of the most common types of structural variation. Although such variants have many biological impacts on species evolution and health, they have been understudied because they are very difficult to detect with short read re-sequencing data. Recently, with the commercialization of novel long reads technologies, insertion variants are finally being discovered and referenced in human populations. Thanks to several international efforts, some gold standard call sets have been produced in 2019, referencing tens of thousands insertions. On these datasets, all existing short-read insertion variant callers, including our own method MindTheGap [9] which overtook others on simulated data, can reach at most 5 to 10 % of the referenced insertion variants. In this work, we propose a classification of the different types of insertion variants, based on the genomic context of the insertion site and the levels of duplication contained in the inserted sequence or within its breakpoints. In a detailed benchmark, we then analyze which of these types are the most impacted by the low recall of existing methods. Finally, by simulating various identified factors of difficulty, we investigate the causes of low recall and how these can be bypassed or improved in existing algorithms.

7.4.3. Modeling activities in cooperation with Inria project Dyliss

Participant: Jacques Nicolas.

J. Nicolas has maintained a partial activity with its previous research team Dyliss. In this framework, we have explored the use of Formal Concept Analysis (FCA) to ease the analysis of biological networks. The PhD thesis of L. Bourneuf on graph compression using FCA, defended this year, has introduced a new extension of FCA for this purpose, working on triplet concepts, which correspond to overlapping bicliques in graphs. The search space of concepts for graph compression has been presented in [21]. FCA applied to data on the steady states of a Boolean network and the dependencies between its proteins allowed to build a classifier used to analyze the states according to the phenotypic signatures of its network components. We have identified variants to the phenotypes and characterized hybrid phenotypes [19].

7.5. Bioinformatics Analysis

7.5.1. Genomics of Brassicaceae and agro-ecosystems insects

Participant: Fabrice Legeai.

Through its long term collaboration with INRA IGEPP, and its support to the BioInformatics of Agroecosystems Arthropods platform (http://bipaa.genouest.org), GenScale is involved in various genomic projects in the field of agricultural research. First, on plant genomics, we helped to identify duplicated copies of genes and repeated elements in the Brassica genomes [14]. Then, on major agricultural pests or their natural ennemies such as parasitoids, we conducted large scale analyses on the expression of effector genes involved in the adaptation of pea aphids to their host-plants [13]. Finally, we explored the expression of genes related to the virus machinery of bathyplectes parasitoids wasp of the alfalfa weevil [18].

7.5.2. Structural genome analysis of S. pyogenes strains

Participants: Emeline Roux, Dominique Lavenier.

The *S. pyogenes* bacteria is responsible for many human infections. With the increase in the prevalence of infections (750 million infections per year worldwide and 4th in terms of mortality from bacterial infection), a better understanding of adaptive and evolutionary mechanisms at play in this bacteria is essential. The molecular characterization of the different strains is done by the *emm* gene. A statistical analysis of the different types of *emm* on the Brittany population shows 3 main dynamics: sporadic types, endemic types or epidemic types. The last case was observed in Brittany for the type *emm75* between 2009 and 2017. Two hypotheses can be considered: (1) the emergence of a new subtype or winning clone in an unimmunized population; (2) increased pathogenicity through genetic evolution of the strains, including the acquisition of new virulence factors. In collaboration with the microbiology department of the Rennes Hospital, we sequenced more than 30 *S. pyogenesemm75* strains (Oxford Nanopore MinION sequencing) in order to study the dynamic of the epidemic thought their structural genomic variation.

7.5.3. Linking allele-specific expression and natural selection in wild populations

Participants: Mohammed Amin Madoui, Pierre Peterlongo.

Allele-specific expression (ASE) is now a widely studied mechanism at cell, tissue and organism levels. However, population-level ASE and its evolutionary impacts have still never been investigated. Here, we hypothesized a potential link between ASE and natural selection on the cosmopolitan copepod *Oithona similis*. We combined metagenomic and metatranscriptomic data from seven wild populations of the marine copepod *O. similis* sampled during the Tara Oceans expedition. We detected 587 single nucleotide variants (SNVs) under ASE and found a significant amount of 152 SNVs under ASE in at least one population and under selection across all the populations. This constitutes a first evidence that selection and ASE target more common loci than expected by chance, raising new questions about the nature of the evolutionary links between the two mechanisms [33].

SERPICO Project-Team

7. New Results

7.1. Empirical SURE-guided microscopy super-resolution image reconstruction from confocal multi-array detectors

Participants: Sylvain Prigent, Charles Kervrann.

Recent confocal microscopes use an array detector instead of single point detector to take multiple views of the same sample. The microscope output is then an array of images, one image per detector. The array of images is then processed to build a single image with higher signal-to-noise ratio and higher resolution than a classical confocal microscope image. In the literature, several methods have been recently proposed to reconstruct the single high resolution image: i/ the ISM method combines array registration and Wiener deconvolution; ii/ the IFED method estimates a high resolution image by subtracting the outer detectors of the array to the inner detectors; iii/ the ISFED consists in subtracting the outer registered detectors and the outer raw images. In that context, we proposed a SURE-guided (Stein's unbiased risk estimation) estimation method to automatically select the parameter ϵ controlling the IFED and ISFED reconstruction algorithms (see Figure 4). We showed on real data that the proposed method is capable to achieve a resolution close to 100 nm without any deconvolution method.

Software: AiryscanJ (see Section 6.4).
Collaborator: S. Dutertre (IGDR – Institute Genetics & Development of Rennes).

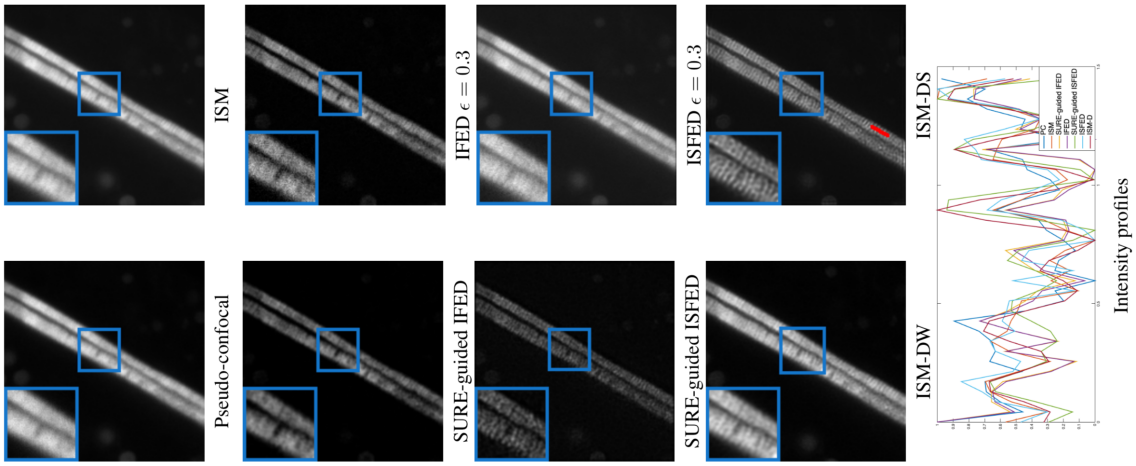


Figure 4. Results obtained on a c.elegans sample and intensity profiles for all the tested methods. The profile line is shown with a red line on the ISM-DS image (top right).

7.2. Dense mapping of intracellular diffusion and drift from single-particle tracking data analysis

Participants: Antoine Salomon, Cesar Augusto Valades Cruz, Charles Kervrann.

It is of primary interest for biologists to be able to locally estimate diffusion and drift inside a cell. In our framework, we assumed that particle motion is governed by the following Langevin equation: $\mathbf{dx} = \mathbf{b}(\mathbf{x})dt + \sigma(\mathbf{x})\mathbf{dw}$ where $\mathbf{x}$ denotes the 2D or 3D spatial coordinates of the particle, $\mathbf{b}$ the drift vector, σ the diffusion coefficient, and $\mathbf{w}$ the standard Gaussian white noise. In that context, we proposed a new mapping method inspired from [50] that developed a method providing results in the form of matrices by scanning the data by blocks, and from the framework in [5], dedicated to both classifying particle motion types and detecting potential motion switches along a trajectory. To avoid the calculation of both drift and diffusion in cell coordinates where no data is available, we replaced the scanning movement of an averaging window by a Gaussian window centered on trajectory points. Each drift vector and each diffusion coefficient are calculated at coordinates corresponding exactly to the coordinates given by the preliminary particle tracking, which provides more details. A nonparametric three-decision test enables to label trajectories or sub-trajectories [5]. This information is then used. to calculate drift and diffusion coefficient (or Kramers-Moyal coefficients) maps separately on each class of motion with the most appropriate diffusion models: confined motion (the particle is bound to a specific point), Brownian motion (the particle moves randomly), and directed motion (the particle moves in a given direction) (see Figure 5).

Software: THOTH and CPAnalysis (see Sections 6.2 and 6.6).

Collaborators: V. Briane (UNSW Sydney, School of Medical Sciences, Australia),
L. Leconte and J. Salamero (CNRS-UMR 144, Institut Curie, PSL Research University),
L. Johannes (U1143 INSERM / CNRS-UMR 3666, Institut Curie, PSL Research University),
E. Derivery (MRC laboratory of Molecular Biology, Cambridge, UK),
L. Muresan (Cambridge Advanced Imaging Centre, UK).

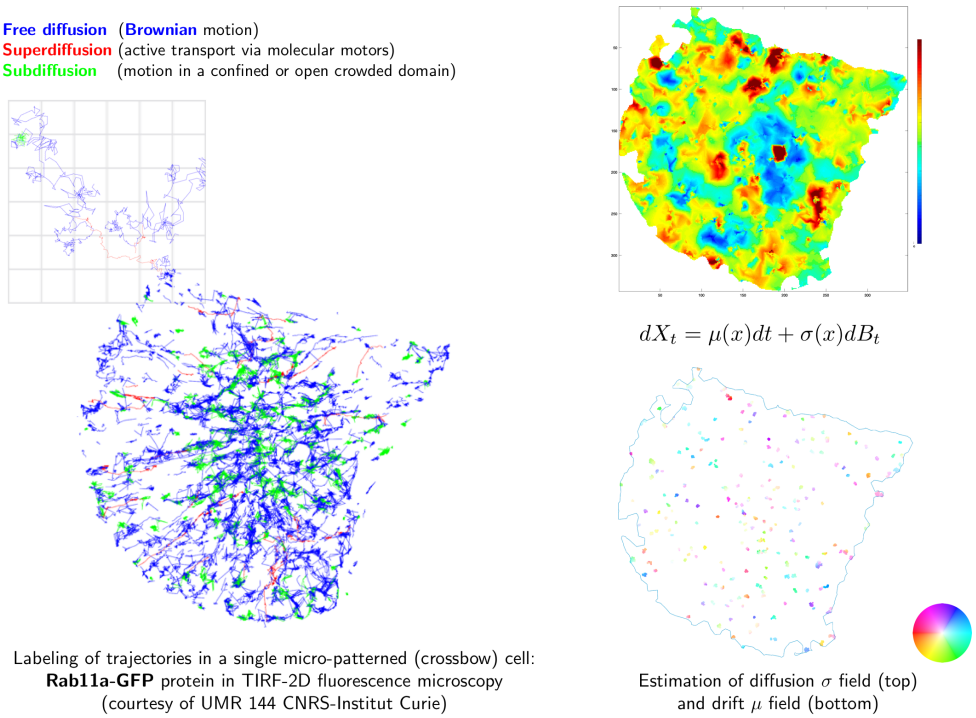


Figure 5. Intracellular diffusion and drift maps estimated from simulated tracking data (FluoSIM).

7.3. 3D tracking of endocytic and exocytic events using lattice light sheet microscopy

Participants: Cesar Augusto Valades Cruz, Ludovic Leconte, Jean Salamero, Charles Kervrann.

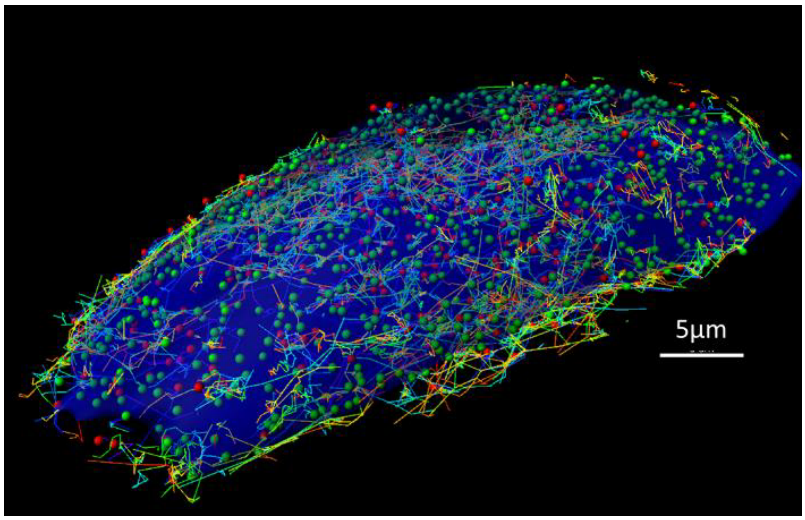


Figure 6. 3D tracking of Gal3-Atto647n (red) vs AP2-eGFP (green) adaptor protein in SUM159 cell.

The study of the whole cell dynamics of endocytic/exocytic-recycling events has proven difficult until recently because of lack of sensitivity, limited speed, photobleaching and phototoxicity associated with conventional imaging modalities. The Lattice Light Sheet Microscope (LLSM) allows to overcome these difficulties, yet reaching high spatial resolution. 3D images are captured for several minutes at a high acquisition frequency, and enables the study of signaling, transport, and stochastic self-assembly in complex environments. In addition, this imaging technique and 3D-tracking allow us to characterize the molecular machineries involved in the exocytosis and endocytosis mechanisms. We have the opportunity to observe a series of sequential events corresponding to the fusion with the plasma membrane (exocytosis) and the formation of endocytic carriers, including the trafficking of vesicles throughout the entire membrane system. We have got preliminary results of the coordination of vesicle recycling from the endosomal recycling compartment up to the plasma membrane using LLSM imaging and 3D tracking. In addition, we introduced a quantitative analysis of endocytosis dynamics of AP2 adaptor complex, Galectin-3 (see Figure 6) and Transferrin using single particle tracking analysis of 3D+time data. These case studies clearly demonstrated the advantage of lattice light sheet microscopy for imaging endocytic/exocytic events in single cells.

Software: THOTH and CPAnalysis (see Sections 6.2 and 6.6).

Collaborators: C. Wunder and L. Johannes (Institut Curie, PSL Research University, Cellular and Chemical Biology, U1143 INSERM / UMR 3666 CNRS).

7.4. 3D flow estimation in 3D fluorescence microscopy image sequences

Participants: Sandeep Manandhar, Patrick Bouthemy, Charles Kervrann.

We have proposed a variational approach for 3D optical flow computation from a pair of fluorescence microscopy volumes. This computational method has been extensively evaluated on appropriate simulated data. To simulate a volume pair with a ground truth flow field, we extended the Horn-and-Schunck optical flow method to 3D (3DHS). The computed flow field by 3DHS between two input images is then used to generate a new pair volume. The new source and target volumes along with the corresponding 3DHS flow fields serve as the dataset with the ground truth. The latter was used for the parameterization of our method and for the comparative study with the state-of-the art method proposed by Amat et al., 2012 [44].

Meanwhile, we proposed a novel error measure named SAE (for Structural Angular Error) for 3D optical flow, in absence of any ground truth flow field (see Figure 7). SAE measures the angular difference between the principal orientations of the structures present in the backward warped target and the source volumes at each voxel. We found out that the average of SAE (ASAE) and average of the end-point-error (a standard optical flow error in presence of ground truth) behave similarly.

We also integrated L_1 regularization in our variational approach. In contrast to our previous L_2 regularization approach, this method preserves discontinuities of the flow field. However, both of our methods are time demanding and not parallelizable in implementation. Then, we integrated our Census Signature-based data term with total variation regularization that also produces discontinuous flow fields. Consequently, we took the splitting approach for optimization. A gain of four times was obtained in the calculation speed with 12-core implementation of the new method, compared to our previous two methods, for still similar ASAE score.

We have also proposed two new methods for the visualization of the 3D flow fields, named 3DHSV and MIP-flow respectively. The 3DHSV method color codes the 2D projections of the flow field in the Hue and the Saturation color spaces, while mapping the off-the-plane motion to the Value space. MIP-flow also encodes the 2D projections to the Hue and the Saturation spaces. However, it only considers encoding the 2D projections of the 3D vectors corresponding to the maximum intensity points in the direction perpendicular to the projection plane. This work is carried out in collaboration with UTSW Dallas in the frame of the Inria associated team CytoDI.

Software: FlowScope (see Section 6.7).
Collaborators: P. Roudot, E. Welf and G. Danuser (UTSW Medical Center, Dallas, USA).

7.5. Probabilistic overall reconstruction of membrane-associated molecular dynamics from partial observations in rod-shaped bacteria

Participants: Yunjiao Lu, Charles Kervrann.

Understanding the mechanisms that maintain the structure of rod-shaped bacteria is a challenging problem in cell biological research. Thanks to progress in molecular biology and microscopy (e.g Total Internal Reflection Fluorescence (TIRF) microscopy), we have the opportunity to observe the dynamics of the cell wall construction workers, that is the membrane-associated molecular machines (MMs). Due to the cylindrical form of the bacteria and the 2D selective visualization in TIRF microscopy, only around one third of the perimeter can be observed at a given time. Nevertheless, from the partial observed bacteria surface images, earlier studies showed that a fraction of the MMs performs directed motion, across the image field quasi-orthogonally to the cylinder axis.

Accordingly, we addressed the problem of the connection of motion segments on a cylindrical surface, assuming that one MM may re-enters into the observed region (OR), a certain period of time after having left the field of view. The directed MM motions are assumed as Brownian motion with drift. The birth and death events of the MMs are supposed to happen independently and uniformly on the surface. Given a set of observed segments entering and exiting the OR, we proposed a probabilistic framework to calculate the probabilities of the events of birth, death and re-entry, based on speed and diffusion of the motion and the time of exit and entry. Even though two third of the surface is hidden as shown in Figure 8 , this framework allows us to derive a computational procedure aiming at connecting segments belonging to the same trajectory, and then recovering directed MMs dynamics on the whole surface. The performance of the method has been demonstrated on appropriate simulation data that mimics MMs dynamics observed in TIRF microscopy.

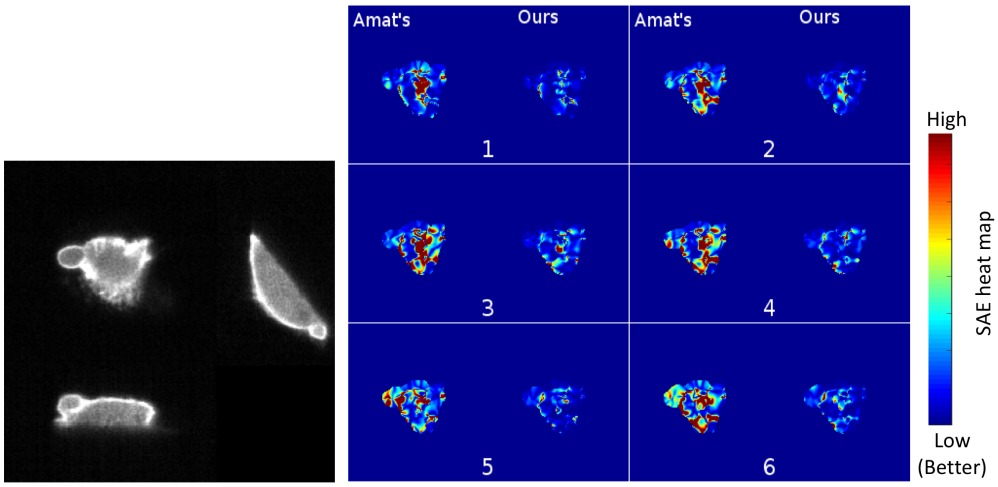


Figure 7. SAE maps (right) to compare our variational method to the Amat's method [44] applied to a 3D image pair (left) depicting a MV3 melanoma cell undergoing blebbing on a coverslip observed with Diagonally scanned Light-Sheet Microscopy (2.86 Hz sampling frequency, $300 \times 300 \times 83/50 \times 50 \times 30\mu m^3$*) (input images by courtesy of Danuser lab, UTSW Dallas, USA).*

Collaborators: A. Trubuil and P. Hodara (INRA MaIAGE unit, Jouy-en-Josas),
R. Carballido-López and C. Billaudeau (INRA, UR MICALIS, Jouy-en-Josas).

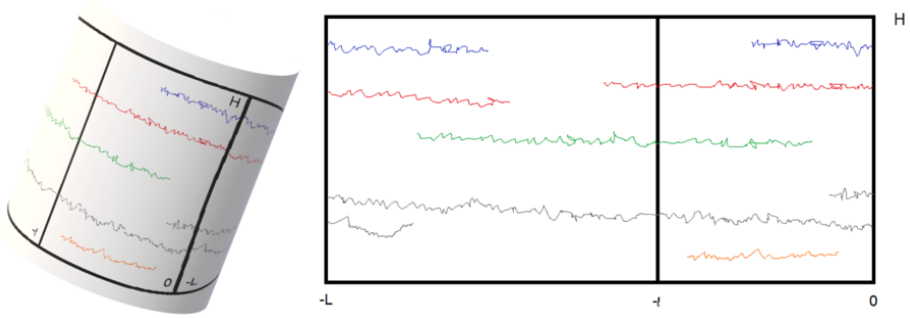


Figure 8. Trajectories on the cylinder and its 2D representation. The unobserved region is $(-L, -l] \times [0, H]$ *and the observed region is* $(l, 0] \times [0, H]$.

7.6. Data assimilation and modeling of cell division mechanism

Participants: Anca-Georgiana Caranfil, Charles Kervrann.

Asymmetric cell division is a complex process that is not yet fully understood. A very well-known example of such a division is the first division of C.elegans embryo. To improve our understanding of this process, we used mathematical modeling to study the first division of C.elegans embryo, both on wild type cells and under a wide range of genetic perturbations. Asymmetry is clearly visible at the end of the anaphase, as the mitotic spindle is off-center. The study of the mitotic spindle dynamics is, thus, a useful tool to gain insights into the general mechanics of the system used by the cell to correctly achieve asymmetric division. The overall spindle behavior is led by the spindle poles behavior. We proposed a new dynamic model for the posterior spindle pole that explains the oscillatory behavior during anaphase and confirms some previous findings, such as the existence of a threshold number of active force-generator motors required for the onset of oscillations. We also confirmed that the monotonic increase of motor activity accounts for their build-up and die-down. By theoretically analyzing our model, we determined boundaries for the motor activity-related parameters for these oscillations to happen. This also allowed us to describe the influence of the number of motors, as well as physical parameters related to viscosity or string-like forces, on features such as the amplitude and number of oscillations. Lastly, by using a Bayesian approach to confront our model to experimental data, we were able to estimate distributions for our biological and bio-physical parameters. These results give us insights on variations in spindle behavior during anaphase in asymmetric division, and provide means of prediction for phenotypes related to misguided asymmetric division. This model will be instrumental in probing the function of yet undocumented genes involved in controlling cell division dynamics.

Collaborators: Y. Le Cunff and J. Pécréaux (IGDR – Institute of Genetics & Development of Rennes).

7.7. Convolutional Neural Networks algorithms for calcium signal segmentation in astrocytes in 3D lattice light sheet microscopy

Participants: Anais Badoual, Charles Kervrann.

Astrocytes, glial cells of the central nervous system, are detectors and regulators of neuronal information processing. It is established that neuronal synapses are physical sites of intercellular contact that transmit and transform information in a very rapid and flexible way, playing a pivotal role for learning and memory formation as well as neurological diseases of the mammalian brain. Astrocytes are thought to integrate neuronal inputs and modulate information transfer between neurons. In particular, cytoplasmic calcium signaling in astrocytes is believed to be crucial for astrocyte-neuron communication. However, quantification of intracellular calcium signals in astrocytes is hindered by the complexity of their cell shape, that consists of a cell body sprouting a highly ramified set of large to very fine protrusions called processes. Until recently, the quantification of intracellular propagation of calcium signal in astrocytes with fluorescent calcium indicators has been restricted to two dimensions, either 2D cell cultures or 2D slicing of a 3D setup. However it is not clear what amount of information is lost by ignoring the 3rd dimension in these experiments. The emergent 3D Lattice Light Sheet Microscopy (LLSM) is a powerful and promising technology (voxel size: 250nm x 250nm x 700nm; acquisition time: 200 frames per second) to give a much more complete and refined view of the dynamic behavior of calcium signaling in astrocytes inside living brain slices and in the intact mouse brain in vivo. Unfortunately, we lack image analysis tools to locate, segment, track and quantify the propagation of those 3D calcium signals in very ramified cell shapes.

In this context, we have started to develop an image processing tool for neurobiologists that 1) detects and segments calcium signals in 3D+time LLSM images, and 2) classifies these signals based on their 3D space-time morphological characterization. To do so, we focus on 3D convolutional network and machine learning techniques.

Collaborators: V. Nägerl and M. Arizono (Interdisciplinary Institute for Neuroscience, Bordeaux), H. Berry and A. Denisot (EPC BEAGLE, Inria Rhone-Alpes).

7.8. Geo-colocalization and coorientation in fluorescence super-resolution microscopy

Participants: Frédéric Lavancier, Reda Alami Chantoufi, Aymeric Lechevranton, Antoine Salomon, Charles Kervrann.

Colocalization aims at characterizing spatial associations between two fluorescently-tagged biomolecules by quantifying the co-occurrence and correlation between the two channels acquired in fluorescence microscopy. This problem remains an open issue in diffraction-limited microscopy and raises new challenges with the emergence of super-resolution imaging. In [19], we proposed an original method (GcoPS) that exploits the random sets structure of the tagged molecules to provide an explicit testing procedure. GcoPS requires the adjustment of a p-value that guarantees more reproducibility and more objective interpretation and takes as inputs two 2D or 3D binary segmented images. This year, we extended this approach to the estimation of local co-localization. This amounts to applying the statistical test on windows randomly drawn in the whole image. A multiple testing procedure allows us to compute a global partial colocalization score. Meanwhile, the excursion sets of colocalization score map estimated by Gaussian smoothing are very helpful to detect regions of interest corresponding to significant colocalization and anti-colocalization sites. This approach has been evaluated on STORM (Stochastic Optical Reconstruction Microscopy) images which provides several hundreds thousands of super-localized positions of individual molecules with an average accuracy of 10-20 nanometers (see Figure 9). Finally, the method has been successively extended to the geo-coorientation (or geo-coalignment) of 2D-3D vectors (optical flow, tensors) and trajectories to analyze the molecular interactions.

Software: GcoPS (see Section 6.1).

Collaborators: J. Salamero (CNRS-UMR 144, Institut Curie, PSL Research University),
G. Bertolin (IGDR – Institute of Genetics & Development of Rennes),
M. Lelek and C. Zimmer (Institut Pasteur, Paris).

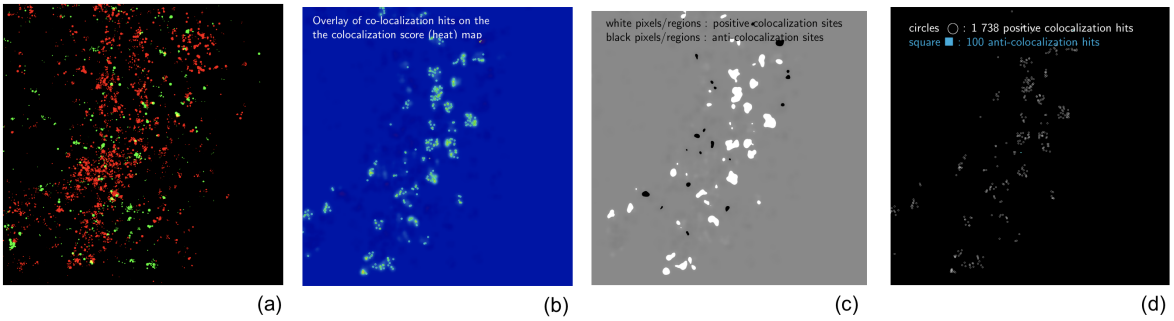


Figure 9. Illustration of geo-colocalization of two molecules (red/green channels) in STORM super-resolution microscopy (original image size: 4576 × 3564 pixels; pixel size: 3 nanometers). (a) Overlay of two channels (sub-region of the original pair); (b) colocalization hits overlaid on the score (heat) map; (c) excursion sets of detected colocalization (white) and anti-colocalization (black) sites overlaid on the score map; (d) detected co-localization (circles) and anti-co-localization (squares) hits.

7.9. Immersive and interactive visualization of 3D temporal data using a space time hypercube

Participants: Gwendal Fouché, Charles Kervrann.

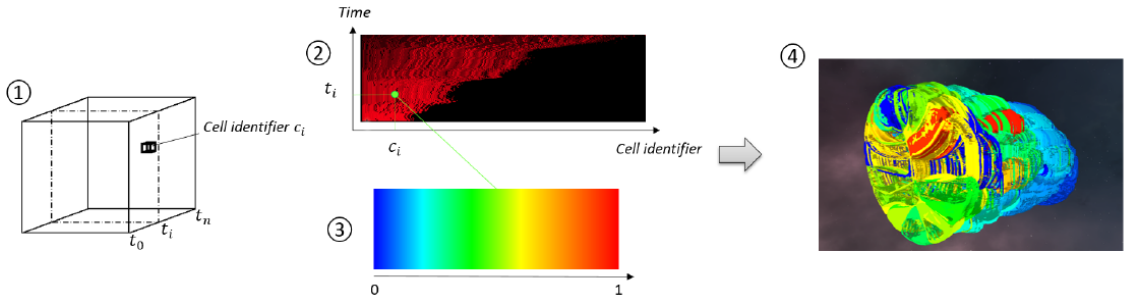


Figure 10. Flow diagram of the STC generation. In step 1, the user places the interactive clipping plan to get the desired cross-section. In step 2, camera parameters are automatically set in order to render the cross section at each time point. In step 3, the top image presents the output of the rendering operation, using the RGB channel to save cell identifiers; the bottom image is the result of an edge detection filter that will be useful for display. In step 4, the rendered images are stacked into a 3D texture. Each voxel contains a cell identifier, and its position in terms of depth indicates a time point t_i.

The analysis of multidimensional time-varying datasets, whose size grows as recording and simulating techniques advance, faces challenges on the representation and visualization of dense data, as well as on the study of temporal variations. In this context, we proposed an extension of the well-known Space-Time Cube (STC) visualization technique in order to visualize time-varying 3D spatial data acquired in 3D fluorescence microscopy, taking advantage of the interaction capabilities of Virtual Reality. The extended STC is based on a user-driven projection of the spatial and temporal information modeled as a 4D Space-Time Hypercube (STH). This projection yields a 3D STC visualization, which can also encode non-spatial quantitative data. Moreover, we proposed a set of tools allowing the user to manipulate the 3D STC that benefits from the visualization, exploration and interaction possibilities offered by immersive environments (see Figure 10). Finally, the extended STC has been integrated in a VR application for visualization of spatiotemporal biological data, illustrating the usage of the proposed visualization method for the morphogenesis analysis.

Collaborators: F. Argelaguet (EPC HYBRID, Inria Rennes),
E. Faure (Laboratory of Computer Science, Robotics and Microelectronics of Montpellier).

7.10. Unsupervised motion saliency map estimation based on optical flow inpainting

Participants: Léo Maczyta, Patrick Bouthemy.

We have addressed the problem of motion saliency in videos. Salient moving regions are regions that exhibit motion departing from their spatial context in the image, that is, different from the surrounding motion. In contrast to video saliency approaches, we estimate dynamic saliency based on motion information only. We propose a new unsupervised paradigm to compute motion saliency maps. The key ingredient is the flow inpainting stage. We have to compare the flow field in a given area, likely to be a salient moving element, with the flow field that would have been induced in the very same area with the surrounding motion. The former can be computed by any optical flow method. The latter is not directly available, since it is not observed. Yet, it can be predicted by a flow inpainting method. This is precisely the originality of our motion saliency approach.

Our method is then two-fold. First, we extract candidate salient regions from the optical flow boundaries. Secondly, we estimate the inpainted flow using an extension of a diffusion-based method for image inpainting, and we compare the inpainted flow to the original optical flow in these regions. We interpret the possible discrepancy (or residual flow) between the two flows as an indicator of motion saliency. In addition, we combine a backward and forward processing of the video sequence. The method is flexible and general enough, by relying on motion information only. Experimental results on the DAVIS 2016 benchmark demonstrate that the method compares favorably with state-of-the-art video saliency methods. Additionally, by estimating the residual flow, we provide additional information regarding motion saliency that could be further exploited (see Figure 11).

Collaborators: O. Le Meur (Percept team, IRISA, Rennes).

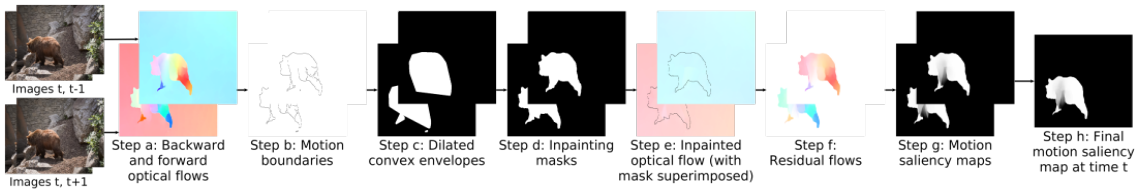


Figure 11. Workflow of the motion saliency estimation method.

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation

Participants: Gerardo Rubino, Bruno Sericola.

Fluid Queues. Stochastic fluid flow models and, in particular, those driven by Markov chains, have been intensively studied in the last two decades. Not only they have been proven to be efficient tools to mimic Internet traffic flows at a macroscopic level but they are useful tools in many areas of applications such as manufacturing systems or in actuarial sciences, to cite but a few. We propose in [61] a chapter which focus on such a model in the context of performance analysis of a potentially congested system. The latter is modeled by means of a finite-capacity system whose content is described by a Markov driven stable fluid flow. We describe step-by-step a methodology to compute exactly the loss probability of the system. Our approach is based on the computation of hitting probabilities jointly with the peak level reached during a busy period, both in the infinite and finite buffer case. Accordingly we end up with differential Riccati equations that can be solved numerically. Moreover, we are able to characterize the complete distribution of both the duration of congestion and of the total information lost during such a busy period.

Connecting irreducible and absorbing Markov chains. Irreducible Markov chains in continuous time are the basic tool for instance in performance evaluation (typically, a queuing model), where in a large majority of cases, we are interested in the behavior of the modeled system in steady-state. Most metrics used are based on the stationary distribution of the model, under unicity natural conditions. Absorbing Markov chains, also in continuous time, play the equivalent role in dependability evaluation, because realistic models must have a finite lifetime, which corresponds here to the absorption time of the chain. In this case, the object of interest is this lifetime, steady-state gives no useful information about the system, and most of the used metrics are defined based on that object. In [30] with describe different connections between the two worlds together with some consequences of those relations in both areas, that is, both in performance and in dependability.

Transient analysis of Markov queueing models. Analyzing the transient behavior of a queueing system is much harder than studying its steady state, the difference being basically that of moving from a linear system to a linear differential system. However, a huge amount of efforts has been put on the former problem, from all kinds of points of view: trials to find closed-forms of the main state distributions, algorithms for numerical evaluations, approximations of different types, exploration of other transient metrics than the basic state distributions, etc. In [62] we focus on the first two elements, the derivation of closed-forms for the main transient state distributions, and the development of numerical techniques. The chapter is organized as a survey, and the main goal is to position and to underline the role of the uniformization technique, for both finding closed-forms and for developing efficient numerical evaluation procedures. In some cases, we extend the discussion to other related transient metrics that are relevant for applications.

7.2. Distributed Systems

Participants: Hamza Ben Ammar, Yann Busnel, Yassine Hadjadj-Aoul, Yves Mocquard, Frédérique Robin, Bruno Sericola.

Stream Processing Systems. Stream processing systems are today gaining momentum as tools to perform analytics on continuous data streams. Their ability to produce analysis results with sub-second latencies, coupled with their scalability, makes them the preferred choice for many big data companies.

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies.

Membership management is a classic and fundamental problem in many use cases. In networking for instance, it is useful to check if a given IP address belongs to a black list or not, in order to allow access to a given server. This has also become a key issue in very large-scale distributed systems, or in massive databases. Formally, from a subset belonging to a very large universe, the problem consists in answering the question "Given any element of the universe, does it belong to a given subset?". Since the access of a perfect oracle answering the question is commonly admitted to be very costly, it is necessary to provide efficient and inexpensive techniques in the context where the elements arrive continuously in a data stream (for example, in network metrology, log analysis, continuous queries in massive databases, etc.). In [36], we propose a simple but efficient solution to answer membership queries based on a couple of Bloom filters. In a nutshell, the idea is to contact the oracle only if an item is seen for the first time. We use a classical Bloom filter to remember an item occurrence. For the next occurrences, we answer the membership query using a second Bloom filter, which is dynamically populated only when the database is queried. We provide theoretical bounds on the false positive and negative probabilities and we illustrate through extensive simulations the efficiency of our solution, in comparison with standard solutions such as classic Bloom filters.

Shuffle grouping is a technique used by stream processing frameworks to share input load among parallel instances of stateless operators. With shuffle grouping each tuple of a stream can be assigned to any available operator instance, independently from any previous assignment. A common approach to implement shuffle grouping is to adopt a Round-Robin policy, a simple solution that fares well as long as the tuple execution time is almost the same for all the tuples. However, such an assumption rarely holds in real cases where execution time strongly depends on tuple content. As a consequence, parallel stateless operators within stream processing applications may experience unpredictable unbalance that, in the end, causes undesirable increase in tuple completion times. We consider recently an application to continuous queries, which are processed by a stream processing engine (SPE) to generate timely results given the ephemeral input data. Variations of input data streams, in terms of both volume and distribution of values, have a large impact on computational resource requirements. Dynamic and Automatic Balanced Scaling for Storm (DABS-Storm) [21] is an original solution for handling dynamic adaptation of continuous queries processing according to evolution of input stream properties, while controlling the system stability. Both fluctuations in data volume and distribution of values within data streams are handled by DABS-Storm to adjust the resources usage that best meets processing needs. To achieve this goal, the DABS-Storm holistic approach combines a proactive auto-parallelization algorithm with a latency-aware load balancing strategy.

Sampling techniques constitute a classical method for detection in large-scale data streams. We have proposed a new algorithm that detects on the fly the k most frequent items in the sliding window model [52]. This algorithm is distributed among the nodes of the system. It is inspired by a recent approach, which consists in associating a stochastic value correlated with the item's frequency instead of trying to estimate its number of occurrences. This stochastic value corresponds to the number of consecutive heads in coin flipping until the first tail occurs. The original approach was to retain just the maximum of consecutive heads obtained by an item, since an item that often occurs will have a higher probability of having a high value. While effective for very skewed data distributions, the correlation is not tight enough to robustly distinguish items with comparable frequencies. To address this important issue, we propose to combine the stochastic approach with a deterministic counting of items. Specifically, in place of keeping the maximum number of consecutive heads obtained by an item, we count the number of times the coin flipping process of an item has exceeded a given threshold. This threshold is defined by combining theoretical results in leader election and coupon collector problems. Results on simulated data show how impressive is the detection of the top-k items in a large range of distributions.

Health Big Data Analysis. The aim of the study was to build a proof-of-concept demonstrating that big data technology could improve drug safety monitoring in a hospital and could help pharmacovigilance professionals to make data-driven targeted hypotheses on adverse drug events (ADEs) due to drug-drug interactions (DDI). In [17], we developed a DDI automatic detection system based on treatment data and laboratory tests from the electronic health records stored in the clinical data warehouse of Rennes academic hospital. We also used OrientDb, a graph database to store informations from five drug knowledge databases and Spark to perform analysis of potential interactions between drugs taken by hospitalized patients. Then, we

developed a Machine Learning model to identify the patients in whom an ADE might have occurred because of a DDI. The DDI detection system worked efficiently and the computation time was manageable. The system could be routinely employed for monitoring.

Probabilistic analysis of population protocols. The computational model of population protocols is a formalism that allows the analysis of properties emerging from simple and pairwise interactions among a very large number of anonymous finite-state agents. In [23] we studied dissemination of information in large scale distributed networks through pairwise interactions. This problem, originally called rumor mongering, and then rumor spreading, has mainly been investigated in the synchronous model. This model relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In the paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where at random times, nodes successively interact by pairs exchanging their information on the rumor. In a previous paper, we performed a study of the total number of interactions needed for all the nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow us to compare different protocols, we provided a thorough analysis of the distribution of this total number of interactions together with its asymptotic behavior. In this paper we extend this discrete time analysis by solving a conjecture proposed previously and we consider the continuous time case, where a Poisson process is associated to each node to determine the instants at which interactions occur. The rumor spreading time is thus more realistic since it is the real time needed for all the nodes of the network to discover the rumor. Once again, as most of the existing results involve huge constants, we provide tight bound and equivalent of the complementary distribution of the rumor spreading time. We also give the exact asymptotic behavior of the complementary distribution of the rumor spreading time around its expected value when the number of nodes tends to infinity.

Among the different problems addressed in the model of population protocols, average-based problems have been studied for the last few years. In these problems, agents start independently from each other with an initial integer state, and at each interaction with another agent, keep the average of their states as their new state. In [45] and [63], using a well chosen stochastic coupling, we considerably improve upon existing results by providing explicit and tight bounds of the time required to converge to the solution of these problems. We apply these general results to the proportion problem, which consists for each agent to compute the proportion of agents that initially started in one predetermined state, and to the counting population size problem, which aims at estimating the size of the system. Both protocols are uniform, i.e., each agent's local algorithm for computing the outputs, given the inputs, does not require the knowledge of the number of agents. Numerical simulations illustrate our bounds of the convergence time, and show that these bounds are tight in the sense that among extensive simulations, numerous ones fit very well with our bounds.

Organizing both transactions and blocks in a distributed ledger. We propose in [53] a new way to organize both transactions and blocks in a distributed ledger to address the performance issues of permissionless ledgers. In contrast to most of the existing solutions in which the ledger is a chain of blocks extracted from a tree or a graph of chains, we present a distributed ledger whose structure is a balanced directed acyclic graph of blocks. We call this specific graph a SYC-DAG. We show that a SYC-DAG allows us to keep all the remarkable properties of the Bitcoin blockchain in terms of security, immutability, and transparency, while enjoying higher throughput and self-adaptivity to transactions demand. To the best of our knowledge, such a design has never been proposed.

Performance of caching systems. Several studies have focused on improving the performance of caching systems in the context of Content-Centric Networking (CCN). In [16], we propose a fairly generic model of caching systems that can be adapted very easily to represent different caching strategies, even the most advanced ones. Indeed, the proposed model of a single cache, named MACS, which stands for Markov chain-based Approximation of CCN Caching Systems, can be extended to represent an interconnection of caches under different schemes. In order to demonstrate the accuracy of our model, we proposed to derive models of the two most effective techniques in the literature, namely LCD and LRU-K, which may adapt to changing patterns of access.

One of the most important concerns when dealing with the performance of caching systems is the static or dynamic (on-demand) placement of caching resources. This issue is becoming particularly important with the upcoming advent of 5G. In [33] we propose a new technique exploiting the model previously proposed model in [16], in order to achieve the best trade-off between the centralization of resources and their distribution, through an efficient placement of caching resources. To do so, we model the cache resources allocation problem as a multi-objective optimization problem, which is solved using Greedy Randomized Adaptive Search Procedures (GRASP). The obtained results confirm the quality of the outcomes compared to an exhaustive search method and show how a cache allocation solution depends on the network's parameters and on the performance metrics that we want to optimize.

7.3. Machine learning

Participants: Yassine Hadjadj-Aoul, Corentin Hardy, Quang Pham Tran Anh, Gerardo Rubino, Bruno Sericola, Imane Taibi, César Viho.

Distributed deep learning on edge-devices. A recently celebrated type of deep neural network is the Generative Adversarial Network (GAN). GANs are generators of samples from a distribution that has been learned; they are up to now centrally trained from local data on a single location. We question in [37] the performance of training GANs using a spread dataset over a set of distributed machines, following a gossip approach shown to work on standard neural networks. This performance is compared to the federated learning distributed method, that has the drawback of sending model data to a server. We also propose a gossip variant, where GAN components are gossiped independently. Experiments are conducted with Tensorflow with up to 100 emulated machines, on the canonical MNIST dataset. The position of the paper is to provide a first evidence that gossip performances for GAN training are close to the ones of federated learning, while operating in a fully decentralized setup. Second, to highlight that for GANs, the distribution of data on machines is critical (i.e., i.i.d. or not). Third, to illustrate that the gossip variant, despite proposing data diversity to the learning phase, brings only marginal improvements over the classic gossip approach.

This work is a part of the thesis [14].

Deep reinforcement learning for network slicing. Recent achievements in Deep Reinforcement Learning (DRL) have shown the potential of these approaches to solve combinatorial optimization problems. However, the Deep Deterministic Policy Gradient algorithm (DDPG), which is one of the most effective techniques, is not suitable to deal with large-scale discrete action space, which is the case of the Virtual Network Function-Forwarding Graph (VNF-FG) placement. To deal with this problem, we propose several enhancements to improve DDPG efficiency [25][47]. The conventional DDPG generates only one action per iteration; thus, it slowly explores the action space especially in a large action space. Thus, we propose to enhance the exploration by considering multiple noisy actions. In order to avoid getting stuck at a local minimum, we propose to multiply the number of critic (for Q-value) neural networks [25]. In order to improve further the exploration, we propose in [47] an evolutionary algorithm to evolve these neural networks in order to discover better ones.

The techniques presented above are generic and can be applied to a variety of problems. To make them even more effective for network slicing problems, we have also proposed to combine them with a proposed First-Fit heuristic that allows for even more interesting results.

Machine learning for Indoor Outdoor detection. Detecting whether a mobile user is indoor or outdoor is an important issue which significantly impacts user behavior contextualization and mobile network resource management. In [59] we investigate hybrid/semi-supervised Deep Learning-based methods for detecting the environment of an active mobile phone user. They are based on both labeled and unlabeled large real radio data obtained from inside the network and from 3GPP signal measurements. We have empirically evaluated the effectiveness of the semi-supervised learning methods using new real-time radio data, with partial ground truth information, gathered massively from multiple typical and diversified locations (indoor and outdoor) of mobile users. We also presented an analysis of such schemes as compared to the existing supervised classification methods including SVM and Deep Learning [57].

Cognition of user behavior can be seen as an efficient tool for automation of future mobile networks. The work presented in [51] deals with the user behaviour modeling. The model includes the prediction of two main features related to mobile user context: the environment and the mobility. We investigate Deep Learning based methods for simultaneously detecting the environment and the mobility state. We empirically evaluate the effectiveness of the proposed techniques using real-time radio data, which has been massively gathered from multiple diversified situations of mobile users.

Predicting the future Perceived Quality level with PSQA. PSQA is a technology developed by Dionysos during a period of several years, whose aim is quantifying the Quality of Experience (more precisely, the Perceived Quality) of an application or service built on the Internet around the transport of audio or video-audio signals. The main properties of PSQA are the its accuracy (indistinguishable from a subjective testing session), the fact that it is fully automatic, with no reference, and able to operate in real time. PSQA is based on supervised learning (the tool learns from subjective testing panels); once trained and validated, it works with no human intervention. In the PSQA project we selected the Random Neural Network tool for the supervised learning associated tasks, after a comparison with the available techniques at the beginning of the project. In [31] we recall all these elements, including the numerical aspects on the optimization side of the learning process, and then, we focus in the current developments where the goal is to predict the Perceived Quality in the close future. This includes the description of the Reservoir Computing models for time series forecasting, and of a tool we proposed, called Echo State Queueing Network, which is a mix between Reservoir Computing and Random Neural Networks.

7.4. Future networks and architectures

Participants: Yassine Hadjadj-Aoul, Gerardo Rubino, Quang Pham Tran Anh, Anouar Rkhami.

Machine learning for network slicing. Network Function Virtualization (NFV) provides a simple and effective mean to deploy and manage network and telecommunications' services. A typical service can be expressed in the form of a Virtual Network Function-Forwarding Graph (VNF-FG). Allocating a VNF-FG is equivalent to placing VNFs and virtual links onto a given substrate network considering resources and quality of service (QoS) constraints. The deployment of VNF-FGs in large-scale networks, such that QoS measures and deployment cost are optimized, is an emerging challenge. Single-objective VNF-FGs allocation has been addressed in existing literature; however, there is still a lack of studies considering multi-objective VNF-FGs allocation. In addition, it is not trivial to obtain optimal VNF-FGs allocation due to its high computational complexity even in the single-objective case. Genetic algorithms (GAs) have proved their ability in coping with multi-objective optimization problems, thus we propose, in [26], a GA-based scheme to solve multi-objective VNF-FGs allocation problem. The numerical results confirm that the proposed scheme can provide near Pareto-optimal solutions within a short execution time.

In [25], we explore the potential of deep reinforcement learning techniques for the placement of VNF-FGs. However, it turns out that even the most well-known learning technique is ineffective in the context of a large-scale action space. In this respect, we propose approaches to find out feasible solutions while improving significantly the exploration of the action space. The simulation results clearly show the effectiveness of the proposed learning approach for this category of problems. Moreover, thanks to the deep learning process, the performance of the proposed approach is improved over time.

The placement of services, as described above, is extremely complex. The issue is even more complex when it comes to placing a service on several non-cooperative domains, where the network operators hide their infrastructure to other competing domains. In [56], we address these problems by proposing a deep reinforcement learning based VNF-FG embedding approach. The results provide insights into the behaviors of non-cooperative domains. They also show the efficiency of the proposed VNF-FG deployment approach having automatic inter-domain load balancing.

Consistent QoS routing in SDN networks. The Software Defined Networking (SDN) paradigm proposes to decouple the control plane (decision-making process) and the data plane (packet forwarding) to overcome the limitations of traditional network infrastructures, which are known to be difficult to manage, especially

at scale. Although there are previous works focusing on the problem of Quality of Service (QoS) routing in SDN networks, only few solutions have taken into consideration the network consistency, which reflects the adequacy between the decisions made and the decisions that should be taken. Therefore, we propose, in [19], a network architecture that guarantees the consistency of the decisions to be taken in an SDN network. A consistent QoS routing strategy is, then, introduced in a way to avoid any quality degradation of prioritized traffic, while optimizing resources usage. Thus, we proposed a traffic dispersion heuristic in order to achieve this goal. We compared our approach to several existing framework in terms of best-effort flows average throughput, average video bitrate and video Quality of Experience (QoE). The emulations results, which are performed using the Mininet environment, clearly demonstrate the effectiveness of the proposed methodology that outperforms existing frameworks.

Optical networks. In [20] we attack the so called *Capacity Crunch* crisis announced for optical networks infrastructures. This problem refers to the facts that (i) the transmission capacity of an optical fiber is not limitless, (ii) the bandwidth demand continues to increase exponentially and (iii) the limits are getting dangerously close. The cheapest and shortest-term solution is to increase efficiency, because there are several possibilities to do so. This work is a contribution in that direction. We focus on strongly improving the wavelength assignment procedure by moving to an heterogeneous and flexible process, adapting the dimensioning to the individual users' needs in QoS. In the paper we demonstrate that a non-uniform dimensioning strategy and a tighten QoS provision allows to save significant networks capacity, while simultaneously provisioning to each user the QoS established in its Service Level Agreement.

Survivability of internet services is a significant and crucial challenge in designing future optical networks. A robust infrastructure and transmission protocols are needed to handle such a situation so that the users can maintain communication despite the existence of one or more failed components in the network. For this reason, we present in [40] a generalized approach able to tolerate any failure scenario, to the extent the user can still communicate with the remaining components, where a scenario corresponds to an arbitrary set of links in a non-operational state. To assess the survivability problem, we propose a joint solution to the problems listed next. We show how to find a set of primary routes, a set of alternate routes associated with each failure scenario, and the capacity required on the network to allow communication between all users, in spite of the links' failures, while satisfying for each user a specific predefined quality of service threshold, defined in the Service Level Agreement (SLA). Numerical results show that the proposed approach not only enjoys the advantages of low complexity and ease of implementation but is also able to achieve significant resource savings compared to existing methods. The savings are higher than 30% on single link failures and more than a 100% on two simultaneous link failures scenarios as well as in more complex situations.

Network tomography. Internet tomography studies the inference of the internal network performances from end-to-end measurements. For this problem, Unicast probing can be advantageous due to the wide support of unicast and the easy deployment of unicast probing paths. In [48] we propose two statistical generic methods for the inference of additive metrics using unicast probing. Our solutions give more flexibility in the choice of the collection points placement. Moreover, the probed paths are not limited to specific topologies. Firstly, we propose the k-paths method that extends the applicability of a previously proposed solution called Flexicast for tree topologies. It is based on the Expectation-Maximization (EM) algorithm which is characterized by high computational and memory complexities. Secondly, we propose the Evolutionary Sampling Algorithm (ESA) that enhances the accuracy and the computing time but following a different approach. In [49] we present a different approach, targeted at link metrics inference in an SDN/NFV environment (even if it can be exported outside this field) that we called TOM (Tomography for Overlay networks Monitoring). In such an environment, we are particularly interested in supervising network slicing, a recent tool enabling to create multiple virtual networks for different applications and QoS constraints on a Telco infrastructure. The goal is to infer the underlay resources states from the measurements performed in the overlay structure. We model the inference task as a regression problem that we solve following a Neural Network approach. Since getting labeled data for the training phase can be costly, our procedure generates artificial data instead. By creating a large set of random training examples, the Neural Network learns the relations between the measures done at path and link levels. This approach takes advantage of efficient Machine Learning solutions to solve a classic inference problem. Simulations with a public dataset show very promising results compared to statistical-

based methods.We explored mainly additive metrics such as delays or logs of loss rates, but the approach can also be used for non-additive ones such as bandwidth.

7.5. Wireless Networks

Participants: Yann Busnel, Yassine Hadjadj-Aoul, Ali Hodroj, Bruno Sericola, César Viho.

Self-organized UAV-based Supervision and Connectivity. The use of drones has become more widespread in recent years. Many use cases have developed involving these autonomous vehicles, ranging from simple delivery of packages to complex emergency situations following catastrophic events. The miniaturization and very low cost of these machines make it possible today to create large meshes to ensure network coverage in disaster areas, for instance. However, the problems of scaling up and self-organization are still open in these use cases. In [35], we propose a position paper that first presents different new requirements for the deployment of unmanned aerial vehicles (UAV) networks, involving the use of many drones. Then, it introduces solutions from distributed algorithms and real-time data processing to ensure quasi-optimal solutions to the raised problems.

More specifically, providing network services access anytime and anywhere is nowadays a critical issue, especially in disaster emergency situations. A natural response to such a need is the use of autonomous flying drones to help finding survivors and provide network connectivity to the rescue teams. In [34], we propose VESPA, a distributed algorithm using only one-hop information of the drones, to discover targets with unknown location and auto-organize themselves to ensure connectivity between them and the sink in a multi-hop aerial wireless network. We prove that connectivity, termination and coverage are preserved during all stages of our algorithm, and we evaluate the algorithm performances through simulations. Comparison with a prior work shows the efficiency of VESPA both in terms of discovered targets and number of used drones.

Enhancing dynamic adaptive streaming over HTTP for multi-homed users. Mobile video traffic accounted for more than half of all mobile data traffic over the past two years. Due to the limited bandwidth, users demand for high-quality video streaming becomes a challenge, which could be addressed by exploiting the emerging diversity of access network and adaptive video streaming. In [39], a network selection algorithm is proposed for Dynamic Adaptive Streaming over HTTP (DASH),the famous international standard on video streaming, to enhance the received video quality to a "multi-homed user" equipped with multiple interfaces. A Multi-Armed Bandit (MAB) heuristic is proposed for a dynamic selection of the best interface at each step. While the Adaptive Bit rate Rules (ABR) used in DASH allow the video player client to dynamically pick the bit rate level according to the perceived network conditions, at each switching step a quality degradation may occur due to the difference in network conditions of the available interfaces. This paper aims to close this gap by (i) designing a MAB algorithm over DASH for a multi-homed user, (ii) evaluating the proposed mechanism through a test-bed implementation, (iii) extending the classic MAB model and (iv) discussing some open issues.

Vehicular networks. According to recent forecasts, constant population growth and urbanization will bring an additional load of 2.9 billion vehicles to road networks by 2050. This will certainly lead to increased air pollution concerns, highly congested roads putting more strain on an already deteriorated infrastructure, and may increase the risk of accidents on the roads as well. Therefore, to face these issues we need not only to promote the usage of smarter and greener means of transportation but also to design advanced solutions that leverage the capabilities of these means along with modern cities' road infrastructure to maximize its utility. In [38], we explore novel ways of utilizing inter-vehicle and vehicle to infrastructure communication technology to achieve a safe and efficient lane change manoeuvre for Connected and Autonomous Vehicles (CAVs). The need for such new protocols is due to the risk that every lane change manoeuvre brings to drivers and passengers lives in addition to its negative impact on congestion level and resulting air pollution, if not performed at the right time and using the appropriate speed. To avoid this risk, we design two new protocols; one is built upon and extends an existing one, and aims at ensuring a safe and efficient lane change manoeuvre, while the second is an original solution inspired from the mutual exclusion concept used in operating systems. This latter complements the former by exclusively granting lane change permission in a way that avoids any risk of collision.

7.6. Network Economics

Participants: Bruno Tuffin, Patrick Maillé.

The general field of network economics, analyzing the relationships between all acts of the digital economy, has been an important subject for years in the team.

In 2019, we have had a particular focus on network neutrality issues, but trying to look at them from original perspectives, and investigating so-called grey zones not yet addressed in the debate.

What implications of a global Internet with neutral and non-neutral portions? Network neutrality is being discussed worldwide, with different countries applying different policies, some imposing it, others acting against regulation or even repealing it as recently in the USA. The goal of [43] is to model and analyze the interactions of users, content providers, and Internet service providers (ISPs) located in countries with different rules. To do so, we build a simple two-regions game-theoretic model and focus on two scenarios of net neutrality relaxation in one region while it remains enforced in the other one. In a first scenario, from an initial situation where both regions offer the same basic quality, one region allows ISPs to offer fast lanes for a premium while still guaranteeing the basic service; in a second scenario the ISPs in both regions play a game on quality, with only one possible quality in the neutral region, and two in the non-neutral one but with a regulated quality ratio between those. Our numerical experiments lead to very different outcomes, with the first scenario benefiting to all actors (especially the ones in the relaxed-neutrality region) and the second one mainly benefiting mostly to ISPs while Content Providers are worse off, suggesting that regulation should be carefully designed.

Investigating a grey zone: sponsored data. Sponsored data, where content providers have the possibility to pay wireless providers for the data consumed by customers and therefore to exclude it from the data cap, is getting widespread in many countries, but is forbidden in others for concerns of infringing the network neutrality principles. We present in [44] a game-theoretic model analyzing the consequences of sponsored data in presence of competing wireless providers, where sponsoring decided by the content provider can be different at each provider. We also discuss the impact on the proportion of advertising on the displayed content. We show that, surprisingly, the possibility of sponsored data may actually reduce the benefits of content providers and on the other hand increase the revenue of ISPs in competition, with a very limited impact on user welfare.

Search engines, bias, consensus, and search neutrality debate. Different search engines provide different outputs for the same keyword. This may be due to different definitions of relevance, and/or to different knowledge/anticipation of users' preferences, but rankings are also suspected to be biased towards own content, which may prejudicial to other content providers. In [41], we make some initial steps toward a rigorous comparison and analysis of search engines, by proposing a definition for a consensual relevance of a page with respect to a keyword, from a set of search engines. More specifically, we look at the results of several search engines for a sample of keywords, and define for each keyword the visibility of a page based on its ranking over all search engines. This allows to define a score of the search engine for a keyword, and then its average score over all keywords. Based on the pages visibility, we can also define the consensus search engine as the one showing the most visible results for each keyword. We have implemented this model and present an analysis of the results in [41].

7.7. Monte Carlo

Participants: Bruno Tuffin, Gerardo Rubino.

We maintain a research activity in different areas related to dependability, performability and vulnerability analysis of communication systems, using both the Monte Carlo and the Quasi-Monte Carlo approaches to evaluate the relevant metrics. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only tool able to solve complex problems of these types.

Rare event simulation of regenerative systems. Rare events occur by definition with a very small probability but are important to analyze because of potential catastrophic consequences. In [32], we focus on rare event for so-called regenerative processes, that are basically processes such that portions of them are statistically independent of each other. For many complex and/or large models, simulation is the only tool at hand but it requires specific implementations to get an accurate answer in a reasonable time. There are two main families of rare-event simulation techniques: Importance Sampling (IS) and Splitting. In a first part, we briefly remind them and compare their respective advantages but later (somewhat arbitrarily) devote most of the work to IS. We then focus on the estimation of the mean hitting time of a rarely visited set. A natural and direct estimator consists in averaging independent and identically distributed copies of simulated hitting times, but an alternative standard estimator uses the regenerative structure allowing to represent the mean as a ratio of quantities. We see that in the setting of crude simulation, the two estimators are actually asymptotically identical in a rare-event context, but inefficient for different, even if related, reasons: the direct estimator requires a large average computational time of a single run whereas the ratio estimator faces a small probability computation. We then explain that the ratio estimator is advised when using IS. In the third part, we discuss the estimation of the distribution, not just the mean, of the hitting time to a rarely visited set of states. We exploit the property that the distribution of the hitting time divided by its expectation converges weakly to an exponential as the target set probability decreases to zero. The problem then reduces to the extensively studied estimation of the mean described previously. It leads to simple estimators of a quantile and conditional tail expectation of the hitting time. Some variants are presented and the accuracy of the estimators is illustrated on numerical examples.

In [46], we introduce and analyze a new regenerative estimator. A classical simulation estimator of this class is based on a ratio representation of the mean hitting time, using crude simulation to estimate the numerator and importance sampling to handle the denominator, which corresponds to a rare event. But the estimator of the numerator can be inefficient when paths to the set are very long. We thus introduce a new estimator that expresses the numerator as a sum of two terms to be estimated separately. We provide theoretical analysis of a simple example showing that the new estimator can have much better behavior than the classical estimator. Numerical results further illustrate this.

Randomized Quasi-Monte Carlo for Quantile Estimation. Quantile estimation is a key issue in many application domains, but has been proved difficult to efficiently estimate. In [42], we compare two approaches for quantile estimation via randomized quasi-Monte Carlo (RQMC) in an asymptotic setting where the number of randomizations for RQMC grows large but the size of the low-discrepancy point set remains fixed. In the first method, for each randomization, we compute an estimator of the cumulative distribution function (CDF), which is inverted to obtain a quantile estimator, and the overall quantile estimator is the sample average of the quantile estimators across randomizations. The second approach instead computes a single quantile estimator by inverting one CDF estimator across all randomizations. Because quantile estimators are generally biased, the first method leads to an estimator that does not converge to the true quantile as the number of randomizations goes to infinity. In contrast, the second estimator does, and we establish a central limit theorem for it. Numerical results further illustrate these points.

Reliability analysis with dependent components. In the reliability area, the Marshall-Olkin copula model has emerged as the standard tool for capturing dependence between components in failure analysis. In this model, shocks arise at exponential random times, affecting one or several components, thus inducing a natural correlation in the failure process. However, because the number of parameter of the model grows exponentially with the number of components, the tool suffers from the "curse of dimensionality." These models are usually intended to be applied to design a network before its construction; therefore, it is natural to assume that only partial information about failure behavior can be gathered, mostly from similar existing networks. To construct them, we propose in [22] an optimization approach to define the shock's parameters in the copula, in order to match marginal failures probabilities and correlations between these failures. To deal with the exponential number of parameters of the problem, we use a column-generation technique. We also discuss additional criteria that can be incorporated to obtain a suitable model. Our computational experiments show that the resulting tool produces a close estimation of the network reliability, especially when the correlation between component failures is significant.

The Creation Process is an algorithm that transforms a static network model into a dynamic one. It is the basis of different variance reduction methods designed to make efficient reliability estimations on highly reliable networks in which links can only assume two possible values, operational or failed. In [18] the Creation Process is extended to let it operate on network models in which links can assume more than two values. The proposed algorithm, that we called Multi-Level Creation Process, is the basis of a method, also introduced here, to make efficient reliability estimations of highly reliable stochastic flow networks. The method proposed, which consists in an application of Splitting over the Multi-Level Creation Process, is empirically shown to be accurate, efficient, and robust. This work was the first step towards a way to implement an efficient estimation procedure for the problem of flow reliability analysis. Our first solution in that direction was presented in [54], where not only we could develop a procedure providing a significant variance reduction but that allows a direct extension to the final target, the solution to the same estimation problem in the more general case of models where the components are dependent. The idea is an original way of implementing a splitting procedure that leads simultaneously to these two properties.

Rare events in risk analysis. One of the main tasks when dealing with critical systems (systems where specific classes of failures can deal to human losses, or to huge financial losses) is the ability to quantify the associated risks, which is the door that, when opened, leads to paths towards understanding what can happen and why, and towards capturing the relationships existing between the different parts of the system, with respect to those risks. This is also the necessary preliminary work allowing to evaluate the relative importance of different factors, always from the viewpoint of the considered risks, an important component of any disaster management system. Identifying the dominant ones is important to know which parts of the system we must reinforce. The keynote [29] described different tools available for these tasks, and how they can be used depending on the objectives to reach. The focus was on Monte Carlo techniques, the only available ones in general, because the only ones able to evaluate any kind of system, and how they deal with rare events. It also discussed the main related open research problems. The tutorial [64] is closely related to previous talk, but the presentation explores more in general the estimation problem and the main families of techniques available for its solution (Importance Sampling, and the particular case of Zero-Variance methods, Splitting, Recursive Variance Reduction techniques, etc.).

DIVERSE Project-Team

6. New Results

6.1. Results on Variability modeling and management

In general, we are currently exploring the use of machine learning for variability-intensive systems in the context of VaryVary ANR project https://varyvary.github.io.

6.1.1. Variability and testing.

The performance of software systems (such as speed, memory usage, correct identification rate) tends to be an evermore important concern, often nowadays on par with functional correctness for critical systems. Systematically testing these performance concerns is however extremely difficult, in particular because there exists no theory underpinning the evaluation of a performance test suite, i.e., to tell the software developer whether such a test suite is "good enough" or even whether a test suite is better than another one. This work [37] proposes to apply **Multimorphic testing** and empirically assess the effectiveness of performance test suites of software systems coming from various domains. By analogy with mutation testing, our core idea is to leverage the typical configurability of these systems, and to check whether it makes any difference in the outcome of the tests: i.e., are some tests able to "kill" underperforming system configurations? More precisely, we propose a framework for defining and evaluating the coverage of a test suite with respect to a quantitative property of interest. Such properties can be the execution time, the memory usage or the success rate in tasks performed by a software system. This framework can be used to assess whether a new test case is worth adding to a test suite or to select an optimal test suite with respect to a property of interest. We evaluate several aspects of our proposal through 3 empirical studies carried out in different fields: object tracking in videos, object recognition in images, and code generators.

6.1.2. Variability, sampling, and SAT.

Uniform or near-uniform generation of solutions for large satisfiability formulas is a problem of theoretical and practical interest for the testing community. Recent works proposed two algorithms (namely UniGen and QuickSampler) for reaching a good compromise between execution time and uniformity guarantees, with empirical evidence on SAT benchmarks. In the context of highly-configurable software systems (e.g., Linux), it is unclear whether UniGen and QuickSampler can scale and sample uniform software configurations. We perform a thorough experiment on 128 real-world feature models. We find that UniGen is unable to produce SAT solutions out of such feature models. Furthermore, we show that QuickSampler does not generate uniform samples and that some features are either never part of the sample or too frequently present. Finally, using a case study, we characterize the impacts of these results on the ability to find bugs in a configurable system. Overall, our results suggest that we are not there: more research is needed to explore the cost-effectiveness of uniform sampling when testing large configurable systems. More details [51]. In general, we are investigating sampling algorithms for cost-effectively exploring configuration spaces (see also [63], [67].

6.1.3. Variability and 3D printing.

Configurators rely on logical constraints over parameters to aid users and determine the validity of a configuration. However, for some domains, capturing such configuration knowledge is hard, if not infeasible. This is the case in the 3D printing industry, where parametric 3D object models contain the list of parameters and their value domains, but no explicit constraints. This calls for a complementary approach that learns what configurations are valid based on previous experiences. In this work [41], we report on preliminary experiments showing the capability of state-of-the-art classification algorithms to assist the configuration process. While machine learning holds its promises when it comes to evaluation scores, an in-depth analysis reveals the opportunity to combine the classifiers with constraint solvers.

6.1.4. Variability and video processing.

In an industrial project [24], we addressed the challenge of developing a software-based video generator such that consumers and providers of video processing algorithms can benchmark them on a wide range of video variants. We have designed and developed a variability modeling language, called VM, resulting from the close collaboration with industrial partners during two years. We expose the specific requirements and advanced variability constructs we developed and used to characterize and derive variations of video sequences. The results of our experiments and industrial experience show that our solution is effective to model complex variability information and supports the synthesis of hundreds of realistic video variants. From the software language perspective, we learned that basic variability mechanisms are useful but not enough; attributes and multi-features are of prior importance; meta-information and specific constructs are relevant for scalable and purposeful reasoning over variability models. From the video domain and software perspective, we report on the practical benefits of a variability approach. With more automation and control, practitioners can now envision benchmarking video algorithms over large, diverse, controlled, yet realistic datasets (videos that mimic real recorded videos) – something impossible at the beginning of the project.

6.1.5. Variability and adversarial machine learning

Software product line engineers put a lot of effort to ensure that, through the setting of a large number of possible configuration options, products are acceptable and well-tailored to customers' needs. Unfortunately, options and their mutual interactions create a huge configuration space which is intractable to exhaustively explore. Instead of testing all products, machine learning is increasingly employed to approximate the set of acceptable products out of a small training sample of configurations. Machine learning (ML) techniques can refine a software product line through learned constraints and a priori prevent non-acceptable products to be derived. In this work [53], we use adversarial ML techniques to generate adver- sarial configurations fooling ML classifiers and pinpoint incorrect classifications of products (videos) derived from an industrial video generator. Our attacks yield (up to) a 100% misclassification rate and a drop in accuracy of 5%. We discuss the implications these results have on SPL quality assurance.

6.1.6. Variability, Linux and machine learning

Given a configuration, can humans know in advance the build status, the size, the compilation time, or the boot time of a Linux kernel? Owing to the huge complexity of Linux (there are more than 15000 options with hard constraints and subtle interactions), machines should rather assist contributors and integrators in mastering the configuration space of the kernel. We have developed TuxML https://github.com/TuxML/ an open-source tool based on Docker/Python to massively gather data about thousands of kernel configurations. 200K+ configurations have been automatically built and we show how machine learning can exploit this information to predict properties of unseen Linux configurations, with different use cases (identification of influential/buggy options, finding of small kernels, etc.) The vision is that a continuous understanding of the configuration space is undoubtedly beneficial for the Linux community, yet several technical challenges remain in terms of infrastructure and automation.

Two preprints are available [62] and [49].

A talk has been given at Embedded Linux Conference Europe 2019 (co-located with Open Source Summit 2019) in Lyon about "Learning the Linux Kernel Configuration Space: Results and Challenges" [54].

6.1.7. Variability and machine learning

We gave a tutorial [49] at SPLC 2019 and introduce how machine learning can be used to support activities related to the engineering of configurable systems and software product lines. To the best of our knowledge, this is the first practical tutorial in this trending field. The tutorial is based on a systematic literature review [67] and includes practical tasks (specialization, performance prediction) on real-world systems (VaryLaTeX, x264).

6.2. Results on Software Language Engineering

6.2.1. Software Language Extension Problem

The problem of software language extension and composition drives much of the research in *Software Language Engineering* (SLE). Although various solutions have already been proposed, there is still little understanding of the specific ins and outs of this problem, which hinders the comparison and evaluation of existing solutions. In [34], we introduce the Language Extension Problem as a way to better qualify the scope of the challenges related to language extension and composition. The formulation of the problem is similar to the seminal Expression Problem introduced by Wadler in the late nineties, and lift it from the extensibility of single constructs to the extensibility of groups of constructs, i.e., software languages. We provide a comprehensive definition of the actual constraints when considering language extension, and believe the Language Extension Problem will drive future research in SLE, the same way the original Expression Problem helped to understand the strengths and weaknesses of programming languages and drove much research in programming languages.

6.2.2. A unifying framework for homogeneous model composition

The growing use of models for separating concerns in complex systems has lead to a proliferation of model composition operators. These composition operators have traditionally been defined from scratch following various approaches differing in formality, level of detail, chosen paradigm, and styles. Due to the lack of proper foundations for defining model composition (concepts, abstractions, or frameworks), it is difficult to compare or reuse composition operators. In [33], we stipulate the existence of a unifying framework that reduces all structural composition operators to structural merging, and all composition operators acting on discrete behaviors to event scheduling. We provide convincing evidence of this hypothesis by discussing how structural and behavioral homogeneous model composition operators (i.e., weavers) can be mapped onto this framework. Based on this discussion, we propose a conceptual model of the framework, and identify a set of research challenges, which, if addressed, lead to the realization of this framework to support rigorous and efficient engineering of model composition operators for homogeneous and eventually heterogeneous modeling languages.

6.2.3. Advanced and efficient execution trace management for executable domain-specific modeling languages

Executable Domain-Specific Modeling Languages (xDSMLs) enable the application of early dynamic verification and validation (V&V) techniques for behavioral models. At the core of such techniques, execution traces are used to represent the evolution of models during their execution. In order to construct execution traces for any xDSML, generic trace metamodels can be used. Yet, regarding trace manipulations, generic trace metamodels lack efficiency in time because of their sequential structure, efficiency in memory because they capture superfluous data, and usability because of their conceptual gap with the considered xDSML. Our contribution in [26] is a novel generative approach that defines a multidimensional and domain-specific trace metamodel enabling the construction and manipulation of execution traces for models conforming to a given xDSML. Efficiency in time is improved by providing a variety of navigation paths within traces, while usability and memory are improved by narrowing the scope of trace metamodels to fit the considered xDSML. We evaluated our approach by generating a trace metamodel for fUML and using it for semantic differencing, which is an important V&V technique in the realm of model evolution. Results show a significant performance improvement and simplification of the semantic differencing rules as compared to the usage of a generic trace metamodel.

6.2.4. From DSL specification to interactive computer programming environment

The adoption of Domain-Specific Languages (DSLs) relies on the capacity of language workbenches to automate the development of advanced and customized environments. While DSLs are usually well tailored for the main scenarios, the cost of developing mature tools prevents the ability to develop additional capabilities for alternative scenarios targeting specific tasks (e.g., API testing) or stakeholders (e.g., education). In [47],

we propose an approach to automatically generate interactive computer programming environments from existing specifications of textual interpreted DSLs. The approach provides abstractions to complement the DSL specification, and combines static analysis and language transformations to automate the transformation of the language syntax, the execution state and the execution semantics. We evaluate the approach over a representative set of DSLs, and demonstrate the ability to automatically transform a textual syntax to load partial programs limited to a single statement, and to derive a Read-Eval-Print-Loop (REPL) from the specification of a language interpreter.

6.2.5. Live-UMLRT: A Tool for Live Modeling of UML-RT Models

In the context of Model-driven Development (MDD) models can be executed by interpretation or by the translation of models into existing programming languages, often by code generation. In [42] we present Live-UMLRT, a tool that supports live modeling of UML-RT models when they are executed by code generation. Live-UMLRT is entirely independent of any live programming support offered by the target language. This independence is achieved with the help of a model transformation which equips the model with support for, e.g., debugging and state transfer both of which are required for live modeling. A subsequent code generation then produces a self-reflective program that allows changes to the model elements at runtime (through synchronization of design and runtime models). We have evaluated Live-UMLRT on several use cases. The evaluation shows that (1) code generation, transformation, and state transfer can be carried out with reasonable performance, and (2) our approach can apply model changes to the running execution faster than the standard approach that depends on the live programming support of the target language. A demonstration video: https://youtu.be/6GrR-Y9je7Y.

6.2.6. Applying model-driven engineering to high-performance computing: Experience report, lessons learned, and remaining challenges

In [35], we present a framework for generating optimizing compilers for performance-oriented embedded DSLs (EDSLs). This framework provides facilities to automatically generate the boilerplate code required for building DSL compilers on top of the existing extensible optimizing compilers. We evaluate the practicality of our framework by demonstrating a real-world use-case successfully built with it.

6.2.7. Software languages in the wild (Wikipedia)

Wikipedia is a rich source of information across many knowledge domains. Yet, recovering articles relevant to a specific domain is a difficult problem since such articles may be rare and tend to cover multiple topics. Furthermore, Wikipedia's categories provide an ambiguous classification of articles as they relate to all topics and thus are of limited use. In [46], we develop a new methodology to isolate Wikipedia's articles that describe a specific topic within the scope of relevant categories; the methodology uses super- vised machine learning to retrieve a decision tree classifier based on articles' features (URL patterns, summary text, infoboxes, links from list articles). In a case study, we retrieve 3000+ articles that describe software (computer) languages. Available fragments of ground truths serve as an essential part of the training set to detect relevant articles. The results of the classification are thoroughly evaluated through a survey, in which 31 domain experts participated.

6.3. Results on Heterogeneous and dynamic software architectures

We have selected three main contributions for DIVERSE's research axis #4: one is in the field of runtime management of resources for dynamically adaptive system, one in the field of temporal context model for dynamically adaptive system and a last one to improve the exploration of hidden real-time structures of programming behavior at run time.

6.3.1. Resource-aware models@runtime layer for dynamically adaptive system

In Kevoree, one of the goal is to work on the shipping phases in which we aim at making deployment, and the reconfiguration simple and accessible to a whole development team. This year, we mainly explore two main axes.

In the first one, we try to improve the proposed models that could be used at run time to improve resource usage in two domains: cloud computing [30], [57] and energy [58].

6.3.2. Investigating Machine Learning Algorithms for Modeling SSD I/O Performance for Container-based Virtualization

One of the cornerstones of the cloud provider business is to reduce hardware resources cost by maximizing their utilization. This is done through smartly sharing processor, memory, network and storage, while fully satisfying SLOs negotiated with customers. For the storage part, while SSDs are increasingly deployed in data centers mainly for their performance and energy efficiency, their internal mechanisms may cause a dramatic SLO violation. In effect, we measured that I/O interference may induce a 10x performance drop. We are building a framework based on autonomic computing which aims to achieve intelligent container placement on storage systems by preventing bad I/O interference scenarios. One prerequisite to such a framework is to design SSD performance models that take into account interactions between running processes/containers, the operating system and the SSD. These interactions are complex. In this work [30], we investigate the use of machine learning for building such models in a container based Cloud environment. We have investigated five popular machine learning algorithms along with six different I/O intensive applications and benchmarks. We analyzed the prediction accuracy, the learning curve, the feature importance and the training time of the tested algorithms on four different SSD models. Beyond describing modeling component of our framework, this paper aims to provide insights for cloud providers to implement SLO compliant container placement algorithms on SSDs. Our machine learning-based framework succeeded in modeling I/O interference with a median Normalized Root-Mean-Square Error (NRMSE) of 2.5%.

6.3.3. Cuckoo: Opportunistic MapReduce on Ephemeral and Heterogeneous Cloud Resources

Cloud infrastructures are generally over-provisioned for handling load peaks and node failures. However, the drawback of this approach is that a large portion of data center resources remains unused. In this work [57], we propose a framework that leverages unused resources of data centers, which are ephemeral by nature, to run MapReduce jobs. Our approach allows: i) to run efficiently Hadoop jobs on top of heterogeneous Cloud resources, thanks to our data placement strategy, ii) to predict accurately the volatility of ephemeral resources, thanks to the quantile regression method, and iii) for avoiding the interference between MapReduce jobs and co-resident workloads, thanks to our reactive QoS controller. We have extended Hadoop implementation with our framework and evaluated it with three different data center workloads. The experimental results show that our approach divides Hadoop job execution time by up to 7 when compared to the standard Hadoop implementation. In [44], we presented a demo that leverages unused but volatile Cloud resources to run big data jobs. It is based on a learning algorithm that accurately predicts future availability of resources to automatically scale the ran jobs. We also designed a mechanism that avoids interference between the Big data jobs and co-resident workloads. Our solution is based on Open-Source components such as kubernetes and Apache Spark.

6.3.4. Leveraging cloud unused resources for Big data application while achieving SLA

Companies are more and more inclined to use collaborative cloud resources when their maximum internal capacities are reached in order to minimize their TCO. The downside of using such a collaborative cloud, made of private clouds' unused resources, is that malicious resource providers may sabotage the correct execution of third-party-owned applications due to its uncontrolled nature. In this work [43], we propose an approach that allows sabotage detection in a trustless environment. To do so, we designed a mechanism that (1) builds an application fingerprint considering a large set of resources usage (such as CPU, I/O, memory) in a trusted environment using random forest algorithm, and (2) an online remote fingerprint recognizer that monitors application execution and that makes it possible to detect unexpected application behavior. Our approach has been tested by building the fingerprint of 5 applications on trusted machines. When running these applications on untrusted machines (with either homogeneous, heterogeneous or unspecified hardware from the one that was used to build the model), the fingerprint recognizer was able to ascertain whether the execution of the application is correct or not with a median accuracy of about 98% for heterogeneous hardware and about 40% for the unspecified one.

6.3.5. Benefits of Energy Management Systems on local energy efficiency, an agricultural case study

Energy efficiency is a concern impacting both ecology and economy. Most approaches aiming at reducing the energy impact of a site focus on only one specific aspect of the ecosystem: appliances, local generation or energy storage. A trade-off analysis of the many factors to consider is challenging and must be supported by tools. This work proposes a Model-Driven Engineering approach mixing all these concerns into one comprehensive model [58]. This model can then be used to size either local production means, either energy storage capacity and also help to analyze differences between technologies. It also enables process optimization by modeling activity variability: it takes the weather into account to give regular feedback to the end user. This approach is illustrated by simulation using real consumption and local production data from a representative agricultural site. We show its use by: sizing solar panels, by choosing between battery technologies and specification and by evaluating different demand response scenarios while examining the economic sustainability of these choices.

6.4. Results on Diverse Implementations for Resilience

Diversity is acknowledged as a crucial element for resilience, sustainability and increased wealth in many domains such as sociology, economy and ecology. Yet, despite the large body of theoretical and experimental science that emphasizes the need to conserve high levels of diversity in complex systems, the limited amount of diversity in software-intensive systems is a major issue. This is particularly critical as these systems integrate multiple concerns, are connected to the physical world, run eternally and are open to other services and to users. Here we present our latest observational and technical results about (i) observations of software diversity mainly through browser fingerprinting, and (ii) software testing to study and assess the validity of software.

6.4.1. Privacy and Security

6.4.1.1. A Collaborative Strategy for Mitigating Tracking through Browser Fingerprinting

Browser fingerprinting is a technique that collects information about the browser configuration and the environment in which it is running. This information is so diverse that it can partially or totally identify users online. Over time, several countermeasures have emerged to mitigate tracking through browser fingerprinting. However, these measures do not offer full coverage in terms of privacy protection, as some of them may introduce inconsistencies or unusual behaviors, making these users stand out from the rest. In this work [45], we address these limitations by proposing a novel approach that minimizes both the identifiability of users and the required changes to browser configuration. To this end, we exploit clustering algorithms to identify the devices that are prone to share the same or similar fingerprints and to provide them with a new non-unique fingerprint. We then use this fingerprint to automatically assemble and run web browsers through virtualization within a docker container. Thus all the devices in the same cluster will end up running a web browser with an indistinguishable and consistent fingerprint.

6.4.2. Software Testing

6.4.2.1. A Snowballing Literature Study on Test Amplification

The adoption of agile development approaches has put an increased emphasis on developer testing, resulting in software projects with strong test suites. These suites include a large number of test cases, in which developers embed knowledge about meaningful input data and expected properties in the form of oracles. This work [29] surveys various works that aim at exploiting this knowledge in order to enhance these manually written tests with respect to an engineering goal (e.g., improve coverage of changes or increase the accuracy of fault localization). While these works rely on various techniques and address various goals, we believe they form an emerging and coherent field of research, which we call ‘test amplification’. We devised a first set of papers from DBLP, looking for all papers containing ‘test’ and ‘amplification’ in their title. We reviewed the 70 papers in this set and selected the 4 papers that fit our definition of test amplification. We use these 4 papers as the seed for our snowballing study, and systematically followed the citation graph. This study is the first that draws a comprehensive picture of the different engineering goals proposed in the literature for test amplification. In

particular, we note that the goal of test amplification goes far beyond maximizing coverage only. We believe that this survey will help researchers and practitioners entering this new field to understand more quickly and more deeply the intuitions, concepts and techniques used for test amplification.

6.4.2.2. Automatic Test Improvement with DSpot: a Study with Ten Mature Open-Source Projects

In the literature, there is a rather clear segregation between manually written tests by developers and automatically generated ones. In this work, we explore a third solution: to automatically improve existing test cases written by developers. We present the concept, design, and implementation of a system called DSpot, that takes developer-written test cases as input (Junit tests in Java) and synthesizes improved versions of them as output. Those test improvements are given back to developers as patches or pull requests, that can be directly integrated in the main branch of the test code base. In this work [28], we have evaluated DSpot in a deep, systematic manner over 40 real-world unit test classes from 10 notable and open-source software projects. We have amplified all test methods from those 40 unit test classes. In 26/40 cases, DSpot is able to automatically improve the test under study, by triggering new behaviors and adding new valuable assertions. Next, for ten projects under consideration, we have proposed a test improvement automatically synthesized by DSpot to the lead developers. In total, 13/19 proposed test improvements were accepted by the developers and merged into the main code base. This shows that DSpot is capable of automatically improving unit-tests in real-world, large-scale Java software.

6.4.2.3. Leveraging metamorphic testing to automatically detect inconsistencies in code generator families

Generative software development has paved the way for the creation of multiple code generators that serve as a basis for automatically generating code to different software and hardware platforms. In this context, the software quality becomes highly correlated to the quality of code generators used during software development. Eventual failures may result in a loss of confidence for the developers, who will unlikely continue to use these generators. It is then crucial to verify the correct behaviour of code generators in order to preserve software quality and reliability. In this work [25], we leverage the metamorphic testing approach to automatically detect inconsistencies in code generators via so-called “metamorphic relations”. We define the metamorphic relation (i.e., test oracle) as a comparison between the variations of performance and resource usage of test suites running on different versions of generated code. We rely on statistical methods to find the threshold value from which an unexpected variation is detected. We evaluate our approach by testing a family of code generators with respect to resource usage and performance metrics for five different target software platforms. The experimental results show that our approach is able to detect, among 95 executed test suites, 11 performance and 15 memory usage inconsistencies.

6.4.3. Software Co-evolution

6.4.3.1. An Empirical Study on the Impact of Inconsistency Feedback during Model and Code Co-changing

Model and code co-changing is about the coordinated modification of models and code during evolution. Intermittent inconsistencies are a common occurrence during co-changing. A partial co-change is the period in which the developer changed, say, the model but has not yet propagated the change to the code. Inconsistency feedback can be provided to developers for helping them to complete partial co-changes. However, there is no evidence whether such inconsistency feedback is useful to developers. To investigate this problem, we conducted a controlled experiment with 36 subjects who were required to complete ten partially completed change tasks between models and code of two non-trivial systems [31]. The tasks were of different levels of complexity depending on how many model diagrams they affected. All subjects had to work on all change tasks but sometimes with and sometimes without inconsistency feedback. We then measured differences between task effort and correctness. We found that when subjects were given inconsistency feedback during tasks, they were 268% more likely to complete the co-change correctly compared to when they were not given inconsistency feedback. We also found that when subjects were not given inconsistency feedback, they nearly always failed in completing co-change tasks with high complexity where the partially completed changes were spread across different diagrams in the model. These findings suggest that inconsistency feedback (i.e. detection and repair) should form an integral part of co-changing, regardless of whether the code or the model changes first. Furthermore, these findings suggest that merely having access to changes (as with the given partially completed changes) is insufficient for effective co-changing.

6.4.3.2. Detecting and Exploring Side Effects when Repairing Model Inconsistencies

When software models change, developers often fail in keeping them consistent. Automated support in repairing inconsistencies is widely addressed. Yet, merely enumerating repairs for developers is not enough. A repair can as a side effect cause new unexpected inconsistencies (negative) or even fix other inconsistencies as well (positive). To make matters worse, repairing negative side effects can in turn cause further side effects. Current approaches do not detect and track such side effects in depth, which can increase developers' effort and time spent in repairing inconsistencies. This work [66] presents an automated approach for detecting and tracking the consequences of repairs, i.e. side effects. It recursively explores in depth positive and negative side effects and identifies paths and cycles of repairs. This work further ranks repairs based on side effect knowledge so that developers may quickly find the relevant ones. Our approach and its tool implementation have been empirically assessed on 14 case studies from industry, academia, and GitHub. Results show that both positive and negative side effects occur frequently. A comparison with three versioned models showed the usefulness of our ranking strategy based on side effects. It showed that our approach's top prioritized repairs are those that developers would indeed choose. A controlled experiment with 24 participants further highlights the significant influence of side effects and of our ranking of repairs on developers. Developers who received side effect knowledge chose far more repairs with positive side effects and far less with negative side effects, while being 12.3% faster, in contrast to developers who did not receive side effect knowledge.

6.4.3.3. Supporting A Flexible Grouping Mechanism for Collaborating Engineering Teams

Most engineering tools do not provide much support for collaborating teams and today's engineering knowledge repositories lack flexibility and are limited. Engineering teams have different needs and their team members have different preferences on how and when to collaborate. These needs may depend on the individual work style, the role an engineer has, and the tasks they have to perform within the collaborating group. However, individual collaboration is insufficient and engineers need to collaborate in groups. This work [65] presents a collaboration framework for collaborating groups capable of providing synchronous and asynchronous mode of collaboration. Additionally, our approach enables engineers to mix these collaboration modes to meet the preferences of individual group members. We evaluate the scalability of this framework using four real life large collaboration projects. These projects were found from GitHub and they were under active development by the time of evaluation. We have tested our approach creating groups of different sizes for each project. The results showed that our approach scales to support every case for the groups created. Additionally, we scouted the literature and discovered studies that support the usefulness of different groups with collaboration styles.

6.4.4. Software diversification

6.4.4.1. The Maven Dependency Graph: a Temporal Graph-based Representation of Maven Central

The Maven Central Repository provides an extraordinary source of data to understand complex architecture and evolution phenomena among Java applications. As of September 6, 2018, this repository includes 2.8M artifacts (compiled piece of code implemented in a JVM-based language), each of which is characterized with metadata such as exact version, date of upload and list of dependencies towards other artifacts. Today, one who wants to analyze the complete ecosystem of Maven artifacts and their dependencies faces two key challenges: (i) this is a huge data set; and (ii) dependency relationships among artifacts are not modeled explicitly and cannot be queried. In this work [55], we present the Maven Dependency Graph. This open source data set provides two contributions: a snapshot of the whole Maven Central taken on September 6, 2018, stored in a graph database in which we explicitly model all dependencies; an open source infrastructure to query this huge dataset.

6.4.4.2. The Emergence of Software Diversity in Maven Central

Maven artifacts are immutable: an artifact that is uploaded on Maven Central cannot be removed nor modified. The only way for developers to upgrade their library is to release a new version. Consequently, Maven Central accumulates all the versions of all the libraries that are published there, and applications that declare a dependency towards a library can pick any version. In this work [59], we hypothesize that the immutability of Maven artifacts and the ability to choose any version naturally support the emergence of software diversity

within Maven Central. We analyze 1,487,956 artifacts that represent all the versions of 73,653 libraries. We observe that more than 30% of libraries have multiple versions that are actively used by latest artifacts. In the case of popular libraries, more than 50% of their versions are used. We also observe that more than 17% of libraries have several versions that are significantly more used than the other versions. Our results indicate that the immutability of artifacts in Maven Central does support a sustained level of diversity among versions of libraries in the repository.

EASE Project-Team

6. New Results

6.1. Smart City and ITS

Participants: Indra Ngurah, Christophe Couturier, Rodrigo Silva, Frédéric Weis, Jean-Marie Bonnin [contact].

In the last years, we contributed to the specification of the hybrid (ITS-G5 + Cellular) communication architecture of the French field operation test project SCOOP@F. The proposed solution relies on the MobileIP family of standards and the ISO/ETSI ITS Station architecture we contributed to standardize at IETF and ISO. On this topic our contribution mainly focussed on bringing concepts from the state of the art to real equipments. For the last year of the SCOOP@F part 2 project, we took part to the performance evaluation process by providing a test and validation platform for IP mobility protocols (MobileIP, NEMO) and IPsec cyphering. This platform allows us to identify the performance limits of current implementation of mobility and security protocols. Moreover it spotted implementation incompatibilities between the open source implementations of theses protocols (namely UMIP and StrongSwan) and helped the industrial partners of the project to identify associated risks.

InDiD is the logical follow up of SCOOP@F part 2. This 3.5 years long European project (mid 2019-2023) aims at testing ITS applications on a large scale national deployment of connected vehicles and infrastructure. This version of the project specifically complex use cases (so called day 1.5) and urban application. For the beginning of this project, we proposed several innovative use cases. Our "Backward cartography update" scenario has been selected as a priority candidate for implementation. In line with the collaborative approaches of EASE, we propose to use vehicles' observations to inform other vehicles and/or a cartography server about differences between the digital map and the reality.

We also want to explore the benefits of new capabilities of upcoming communication technologies to enrich the interactions between vehicle and smartphones or wearable devices. We defined an architecture for both localisation and communication with vulnerable users (workers in road and construction works). Short range communications between dangers (maneuvering construction vehicles) and workers rely on the advertisement feature of Bluetooth Low Energy (BLE). This connectionless communication mode enables for easy direct communication between any node in the neighborhood. It is inspired from the ITS-Station communication standard and we aim to integrate our work into future versions of the standards. Another contribution in this project aims at enhancing the localisation precision in harsh conditions. Recent version of radio communication standards (eg. Bluetooth 5.1 or 802.11ax) now integrate intrinsic real time localisation primitives giving information such as Angle of Arrival (AoA), Angle of Departure (AoD) or distance evaluation based on Time of Flight (ToF) measurements. We started to study how to merge this information with other localisation evidence sources and how to structure a collaborative framework to share it with other objects in the environment. This early works opens the doors to many other works in the future.

The development of innovative applications for smart cities has also been made possible by the rise of Internet of Things and especially the deployment of numerous low energy devices. The collection of the huge amount of data produced by all these piece of hardware become a challenge for the communication networks. In smart cities, the mobility of vehicles can be used to collect data produced by connected objects and to deliver them to several applications which are delay tolerant. The Vehicular Delay Tolerant Networks (VDTN) can be utilized for such services. We designed DC4LED (Data Collection for Low Energy Devices): a hierarchical VDTN routing which takes advantage of the specific mobility patterns of the various type of vehicles. It provides a low-cost delivery service for applications that need to gather data generated from the field. The idea is to propose a simple routing scheme where cars, taxis, and buses route data hierarchically in a store-carry-forward mechanism to any of the available Internet Point-of-Presences in the city. We compare using simulation tools the performance of DC4LED routing with two legacy VDTN routing schemes which represent

the extreme ends of VDTN routing spectrum: First-contact and Epidemic routing. It show that DC4LED has much lower network overhead in comparison with the two legacy routing schemes, which is advantageous for its implementation scalability. The DC4LED also maintains comparable data delivery probability and latency to Epidemic routing.

The situational viewing and surveillance in cities is one such category of applications which can benefit from various networking solutions available to transport images or data from installed sensor cameras. We explore how our DC4LED mechanism can be used to for a city-wide image and data collection service. We study the networking performance in terms of increasing image sizes that can be transported with respect to varying vehicular density in city. We focus mainly on two technologies for sensors to vehicles communications: ZigBee and ITS-G5. We show that, surprisingly such very simple mechanism could meet the requirements of multiple services.

6.2. Autonomic Maintenance of Optical Networks

Participant: Jean-Marie Bonnin [contact].

The application of classification techniques based on machine learning approaches to analyze the behavior of network users has interested many researchers in the last years. In a recent work, we have proposed an architecture for optimizing the upstream bandwidth allocation in Passive Optical Network (PON) based on the traffic pattern of each user. Clustering analysis was used in association with an assignment index calculation in order to specify for PON users their upstream data transmission tendency. A dynamic adjustment of Service Level Agreement (SLA) parameters is then performed to maximize the overall customers' satisfaction with the network. In this work, we extend the proposed architecture by adding a prediction module as a complementary to the first classification phase. Grey Model GM(1,1) is used in this context to learn more about the traffic trend of users and improve their assignment. An experimental study is conducted to show the impact of the forecaster and how it can overcome the limits of the initial model.

This work has been done in collaboration with IRISA-OCIF team.

6.3. Location assessment from local observations

Participants: Yoann Maurel, Paul Couderc [contact].

Confidence in location is increasingly important in many applications, in particular for crowd-sensing systems integrating user contributed data/reports, and in augmented reality games. In this context, some users can have an interest in lying about their location, and this assumption has been ignored in several widely used geolocation systems because usually, location is provided by the user's device to enhance the user's experience. Two well known examples of applications vulnerable to location cheating are Pokemon Go and Waze.

Unfortunately, location reporting methods implemented in existing services are weakly protected: it is often possible to lie in simple cases or to emit signals that deceive the more cautious systems. For example, we have experimented simple and successful replay attacks against Google Location using this approach, as shown on Figure 3 .

An interesting idea consists in requiring user devices to prove their location, by forcing a secure interaction with a local resource. This idea has been proposed by several works in the literature; unfortunately, this approach requires ad hoc deployment of specific devices in locations that are to be "provable".

We proposed an alternative solution using passive monitoring of Wi-Fi traffic from existing routers. The principle is to collect beacon timestamp observations (from routers) and other attributes to build a knowledge that requires frequent updates to remains valid, and to use statistical test to validate further observations sent by users. Typically, older data collected by a potential attacker will allow him to guess the current state of the older location for a limited timeframe, while the location validation server will get updates allowing him to determine a probability of cheating request.The main strength is its ability to work on existing Wi- Fi infrastructures, without specific hardware. Although it does not offer absolute proof, it makes attacks much more challenging and is simple to implement.

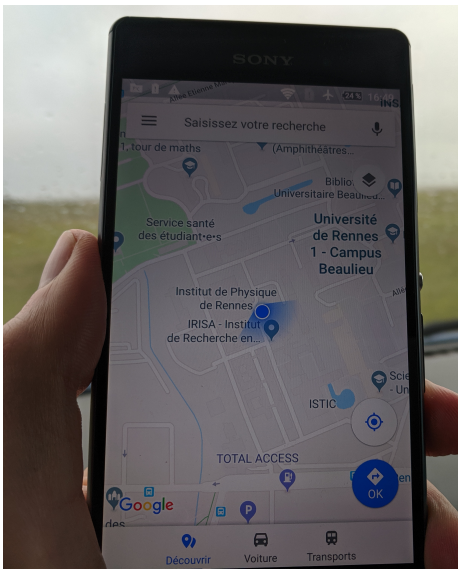


Figure 3. Google map deceived by faked Wi-Fi beacons replayed by an ESP-32

This work was published at CCNC'2019 [1]. We are currently working in broadening this approach, in particular using other attributes of Wi-Fi traffic beside beacon timestamps, and combining the timestamp solution with other type of challenges to propose a diversity of challenges for location validation servers. We are also working on the attack side, which presents interesting perspectives regarding the actual strength of existing services and the potential protection improvements than our approach can provide.

6.4. A methodological framework to promote the use of renewable energy

Participants: Alexandre Rio, Yoann Maurel [contact].

This work is in line with projects aimed at optimizing the use of renewable energies. It is carried out in collaboration with OKWind. This compagny designs and supplies its customers with renewable energy generators such as vertical axis wind turbines and solar trackers. OKWind promotes a micro-grid infrastructure development.

Our application domains are those of agriculture and industry in which it is possible to identify and influence consuming processes. We mainly consider local generation for self-consumption purposes (microgrid) as it limits infrastructure costs, minimizes line losses, reduces the need of the Grid and hopefully reduces the electricity bill.

Renewable energies currently benefit from numerous subsidies to promote their use so as to reduce greenhouse gas emissions. Nevertheless, it seems worth considering the cost-effectiveness of these solutions without these incentives, as they are highly dependent on political will and can be questioned. The reduction in manufacturing costs, particularly in solar energy, suggests that these solutions can eventually compete with traditional sources if they are properly used.

Competitive low-carbon energy is hampered by the stochastic nature of these sources. During peak periods, the electricity produced is competitive, but too often, the scheduled consumption is not aligned with production. In practice, process planning was and is still driven by the electricity price from the grid. On average, the profitability of the installations is therefore not certain. In this context, using battery to shift the load looks appealing but is, as of today, far from being economically viable if not done properly.

Consequently, the achievement of a profitable self-production site is, in practice, a question of trade-off that involves several factors: the scaling of energy sources, the sizing of batteries used, the desired autonomy level, the ecological concerns, and the organization of demand. This trade-off analysis is very challenging: to be carried out effectively and comprehensively, it must be supported by tools that help the stakeholders. While much work has been done in the literature on the impacts of different factors, there are few approaches that offer a comprehensive model.

Our objective is to provide a methodological framework to embrace the diversity of knowledge, of production and consumption tools, of farm activities and of prediction algorithms. This should enable an expert to conduct a trade-off analysis and decide on the best option for each individual site under consideration.

In the first two years of this PhD thesis, we argued that model-driven engineering is suited for the development of such a model and we presented some preliminary implementation. In 2019, we were able to test our approaches in the field and continue to expand the model to account for a wider range of resources. This was published in [5].

6.5. Introducing Data Quality to the Internet of Things

Participants: Jean-Marie Bonnin, Frédéric Weis [contact].

The Internet of Things (IoT) connects various distributed heterogeneous devices. Such Things sense and actuate their physical environment. The IoT pervades more and more into industrial environments forming the so-called Industrial IoT (IIoT). Especially in industrial environments such as smart factories, the quality of data that IoT devices provide is highly relevant. However, current frameworks for managing the IoT and exchanging data do not provide data quality (DQ) metrics. Pervasive applications deployed in the factory need to know how data are "good" for use. However, the DQ requirements differ from a process to another. Actually, specifying/expressing DQ requirements is a subjective task, depending on the specific needs of each targeted application. As an example this could mean how accurate a location of an object that is provided by an IoT system differs from the actual physical position of the object. A Data Quality of 100% could mean that the value represents the actual position. A Data Quality of 0% could mean that the object is not at the reported position. In this example, the value 0% or 100% can be given by a specific software module that is able to filter raw data sent to the IoT system and to deliver the appropriate metric for Dev apps. Building ad hoc solutions for DQ management is perfectly acceptable. But the challenge of writing and deploying applications for the Internet of Things remains often understated. We believe that new approaches are needed, for thinking DQ management in the context of extremely dynamic systems that is the characteristic of the IoT.

In 2019, we introduced DQ to the IoT by (1) representing data quality parameters as metadata to each stored and exchanged IoT data item and (2) providing a toolbox that helps developers to assess the data quality of their processed data using the previously introduced data quality metadata. We followed an inductive approach. Therefore, we set up a pilot to gain first-hand experience with DQ, and to test our developed tools. Our pilot focuses on multi-source data inconsistency. Our setting consists of multiple industrial robots that cowork within a factory. The robots on the line follow a fixed path while the other two robots can freely move. For our implementation we use a data-centric IoT middleware, the Virtual State Layer (VSL). It provides many desired properties such as security and dynamic coupling of services at runtime. Most important it has a strong semantic model for representing data that allows adding new metadata for data quality easily. In our pilot the decrease of the DQ is caused by a low periodicity of location reports. We implemented a DQ service that infers the DQ being located in the service chain. The coordination service queries our DQ enriching service. The DQ enrichment service models the behavior of a robot and infers the resulting DQ depending on the time between the location report and the coordination service's query. Our goal was not only to report the DQ to the consuming service but also to offer tools (microservices) to mitigate from bad DQ. To enable a mitigation from the decreasing DQ, we started the sensors at a random time. This results in the same precision decrease periodicity but in shifted reporting times. The shift enables increasing the DQ by using sensor fusion and data filtering.

KERDATA Project-Team

7. New Results

7.1. Convergence HPC and Big Data

7.1.1. Convergence at the data-processing level

Participants: Gabriel Antoniu, Alexandru Costan, Daniel Rosendo.

Traditional data-driven analytics relies on Big Data processing techniques, consisting of batch processing and real-time (stream) processing, potentially combined in a so-called *Lambda architecture*. This architecture attempts to balance latency, throughput, and fault-tolerance by using batch processing to provide comprehensive and accurate views of batch data, while simultaneously using real-time stream processing to provide views of online data.

On the other side, simulation-driven analytics is based on computational (usually physics-based) simulations of complex phenomena, which often leverage HPC infrastructures. The need to get fast and relevant insights from massive amounts of data generated by extreme-scale simulations led to the emergence of in situ and in transit processing approaches: they allow data to be visualized and processed interactively in real-time as data are produced, while the simulation is running.

To support hybrid analytics and continuous model improvement, we propose to combine the above data processing techniques in what we will call the *Sigma architecture*, a HPC-inspired extension of the Lambda architecture for Big Data processing [17]. Its instantiation in specific application settings depends of course of the specific application requirements and of the constraints that may be induced by the underlying infrastructure. Its main conceptual strength consists in the ability to leverage in a unified, consistent framework, data processing techniques that became reference in HPC in the Big Data communities respectively, without however being combined so far for joint usage in converged environments.

The given framework will integrate previously-validated approaches developed in our team, such as Damaris, a middleware system for efficient I/O management and large-scale in situ data processing, and KerA, a unified system for data flow ingestion and storage. The overall objective is to enable the usage of a large spectrum of Big Data analytics and Intelligence techniques at extreme scales in the Cloud and Edge, to support continuous intelligence (from streaming and historical data) and precise insights/predictions in real-time and fast decision making.

7.1.2. Pufferscale: Elastic storage to support dynamic hybrid workflows systems

Participants: Nathanaël Cheriere, Gabriel Antoniu.

User-space HPC data services are emerging as an appealing alternative to traditional parallel file systems, because of their ability to be tailored to application needs while eliminating unnecessary overheads incurred by POSIX compliance. Such services may need to be rescaled up and down to adapt to changing workloads, in order to optimize resource usage. This can be useful, for instance, to better support complex workflows that mix on-demand simulations and data analytics.

We formalized the operation of rescaling a distributed storage system as a multi objective optimization problem considering three criteria: load balance, data balance, and duration of the rescaling operation. We proposed a heuristic for rapidly finding a good approximate solution, while allowing users to weight the criteria as needed. The heuristic is evaluated with Pufferscale, a new, generic rescaling manager for microservice-based distributed storage systems [18].

To validate our approach in a real-world ecosystem, we showcase the use of Pufferscale as a means to enable storage malleability in the HEPnOS storage system for high energy physics applications.

7.2. Cloud and Edge processing

7.2.1. Benchmarking Edge processing frameworks

Participants: Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

With the spectacular growth of the Internet of Things, edge processing emerged as a relevant means to offload data processing and analytics from centralized Clouds to the devices that serve as data sources (often provided with some processing capabilities). While a large plethora of frameworks for edge processing were recently proposed, the distributed systems community has no clear means today to discriminate between them. Some preliminary surveys exist, focusing on a feature-based comparison.

We claim that a step further is needed, to enable a performance-based comparison. To this purpose, the definition of a benchmark is a necessity. We make a step towards the definition of a methodology for benchmarking Edge processing frameworks [20].

7.2.2. Analytical models for performance evaluation of stream processing

Participants: José Aguilar Canepa, Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

One of the challenges of enabling the Edge computing paradigm is to identify the situations and scenarios in which Edge processing is suitable to be applied. To this end, applications can be modeled as a graph consisting of tasks as nodes and data dependencies between them as edges. The problem comes down to deploying the application graph onto the network graph, that is, operators need to be put on machines, and finding the optimal cut in the graph between the Edge and Cloud resources (i.e., nodes in the network graph).

We have designed an algorithm that finds the optimal execution plan, with a rich cost model that lets users to optimize whichever goal they might be interested in, such as monetary costs, energetic consumption or network traffic, to name a few.

In order to validate the cost model and the effectiveness of the algorithm, a series of experiments were designed using two real-life stream processing applications: a closed-circuit television surveillance system, and an earthquake early warning system.

Two network infrastructures were designed to run the applications. The first one is a state-of-art infrastructure where all processing is done on the Cloud to serve as benchmark. The second one is an infrastructure produced by the algorithm. Both scenarios were executed on the Grid'5000. Several experiments are currently underway. The trade-offs of executing Cloud/Edge workloads with this model were published in [19].

7.2.3. Modeling smart cities applications

Participants: Edgar Romo Montiel, Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

Smart City applications have particular characteristics in terms of data processing and storage, which need to be taken into account by the underlying serving layers. The objective of this new activity is to devise clear models of the data handled by such applications. The data characteristics and the processing requirements does not have to match one-to-one. In some cases, some particular types of data might need one or more types of processing, depending on the use case. For example, small and fast data coming from sensors do not always have to be processed in real-time, but they could also be processed in a batch manner at a later stage.

This activity is the namely the topic of the SmartFastData associated team with the Instituto Politécnico Nacional of Mexico.

In a first phase, we focused on modeling the stream rates of data from sets of sensors in Smart Cities, specifically, from vehicles inside a closed coverage area. Those vehicles are connected in a V2I VANET, and they interact to applications in the Cloud such as traffic reports, navigation apps, multimedia downloading etc. This led to the design of a mathematical model to predict the time that a mobile sensor resides within a geographical designated area.

The proposed model uses Coxian distributions to estimate the time a vehicle requests Cloud services, so that the core challenge is to adjust their parameters. It was achieved by validating the model against real-life data traces from the City of Luxembourg, through extensive experiments on the Grid'5000.

Next, these models were used to estimate the resources needed in the Cloud (or at the Edge) in order to process the whole stream of data. We designed an auto-Scaling module able to adapt the resources with respect to the load. Using the Grid'5000, we evaluated the various possibility to place the prediction module: *(i)* at the Edge, close to data with less accuracy but faster results; or *(ii)* in the Cloud, with higher accuracy due to the global data, but higher latency as well.

7.3. AI across the digital continuum

7.3.1. Machine Learning in the context of Edge stream processing.

Participants: Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

Our research aims to improve the accuracy of Earthquake Early Warning (EEW) systems by means of machine learning. EEW systems are designed to detect and characterize medium and large earthquakes before their damaging effects reach a certain location.

Traditional EEW methods based on seismometers fail to accurately identify large earthquakes due to their sensitivity to the ground motion velocity. The recently introduced high-precision GPS stations, on the other hand, are ineffective to identify medium earthquakes due to its propensity to produce noisy data. In addition, GPS stations and seismometers may be deployed in large numbers across different locations and may produce a significant volume of data consequently, affecting the response time and the robustness of EEW systems.

In practice, EEW can be seen as a typical classification problem in the machine learning field: multi-sensor data are given in input, and earthquake severity is the classification result. We introduce the Distributed Multi-Sensor Earthquake Early Warning (DMSEEW) system, a novel machine learning-based approach that combines data from both types of sensors (GPS stations and seismometers) to detect medium and large earthquakes.

DMSEEW is based on a new stacking ensemble method which has been evaluated on a real-world dataset validated with geoscientists. The system builds on a geographically distributed infrastructure (deployable on clouds and edge systems), ensuring an efficient computation in terms of response time and robustness to partial infrastructure failures. Our experiments show that DMSEEW is more accurate than the traditional seismometer-only approach and the combined-sensors (GPS and seismometers) approach that adopts the rule of relative strength.

These results have been accepted for publication at AAAI, a "A*" conference in the area of Artificial Intelligence [21].

7.3.2. ZettaFlow: Unified Fast Data Storage and Analytics Platform for IoT

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu.

The ZettaFlow platform (system of systems) provides a high-performance multi-model analytics-oriented storage and processing system, while supporting publish-subscribe streams and streaming, key-value and in-memory columnar APIs [16].

The ZettaFlow project is funded by EIT Digital from October 2019 to December 2020. It includes three partners: Inria for the platform development, TU Berlin for edge to cloud IoT optimizations with microservices, and Systematic Paris Region for the go-to-market strategy.

Our goal is to create a startup that will commercialize the ZettaFlow platform: a dynamic, unified and auto-balanced real-time storage and analytics industrial IoT platform. ZettaFlow will provide real-time visibility into machines, assets and factory operations and will automate data driven decisions for high-performance industrial processes.

ZettaFlow will bring a threefold impact to the IoT market.

1. Enable novel real-time edge applications that truly automate manufacturing, transportation and utilities processes.
2. Reduce deployment efforts and time-to-decision of IoT edge-cloud applications by 75% through automation, unified dynamic data management and streaming analytics.
3. Reduce human costs for monitoring and engineering (through edge intelligence) and IoT hardware costs by 50% through unified data collection/storage/analytics.

Myriads Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Efficient Docker container deployment in fog environments

Participants: Arif Ahmed, Lorenzo Civolani, Guillaume Pierre, Paulo Rodrigues de Souza Junior.

Fog computing aims to extend datacenter-based cloud platforms with additional computing, networking and storage resources located in the immediate vicinity of the end users. By bringing computation where the input data was produced and the resulting output data will be consumed, fog computing is expected to support new types of applications which either require very low network latency (e.g., augmented reality applications) or which produce large data volumes which are relevant only locally (e.g., IoT-based data analytics).

Fog computing architectures are fundamentally different from traditional clouds: to provide computing resources in the physical proximity of any end user, fog computing platforms must necessarily rely on very large numbers of small Points-of-Presence connected to each other with commodity networks whereas clouds are typically organized with a handful of extremely powerful data centers connected by dedicated ultra-high-speed networks. This geographical spread also implies that the machines used in any Point-of-Presence may not be datacenter-grade servers but much weaker commodity machines.

We investigated the challenges of efficiently deploying Docker containers in fog platforms composed of tiny single-board computers such as Raspberry PIs. Significant improvements in the Docker image cache hit rate can be obtained by sharing the caches of multiple co-located servers rather than letting them operate independently [9]. In the case when an image must be downloaded and locally installed, large performance gains can be obtained with relatively simple modifications in the way Docker imports container images [3]. Finally, we showed (in collaboration with Prof. Paolo Bellavista from the University of Bologna) that it is possible to let a container start producing useful work even before its image has been fully downloaded [14]. Another paper in this direction of work is in preparation about the way to speedup the boot phase of Docker containers. We are also exploring innovative techniques to improve the performance of live container migration in fog computing environments.

7.1.2. Fog computing platform design

Participants: Ali Fahs, Ayan Mondal, Nikos Parlavantzas, Guillaume Pierre, Mulugeta Tamiru.

There does not yet exist any reference platform for fog computing platforms. We therefore investigated how Kubernetes could be adapted to support the specific needs of fog computing platforms. In particular we focused on the problem of redirecting end-user traffic to a nearby instance of the application. When different users impose various load on the system, any traffic routing system must necessarily implement a tradeoff between proximity and fair load-balancing between the application instances. We demonstrated how such customizeable traffic routing policies can be integrated in Kubernetes to help transform it in a suitable platform for fog computing [15]. We extended this work to let the platform automatically choose (and maintain over time) the best locations where application replicas should be deployed. A paper on this topic is currently under submission. We finally started addressing the topic of application autoscaling such that the system can enforce performance guarantees despite traffic variations. We expect one or two publications on this topic next year.

In collaboration with Prof. Misra from IIT Kharagpur (India), and thanks to the collaboration established by the FogCity associate team, we developed mechanisms based on game theory to assign resources to competing applications in a fog computing platform. The objective of those mechanisms is to satisfy user preferences while maximizing resource utilisation. We evaluated the mechanisms using an emulated fog platform built on Kubernetes and Grid'5000, and showed that they significantly outperform baseline algorithms. A paper on this topic is in preparation.

7.1.3. Edgification of micro-service applications

Participants: Genc Tato, Cédric Tedeschi, Marin Bertier.

Last year, we investigated in collaboration with Etienne Riviere from UC Louvain the feasibility and possible benefits brought about by the *edgification* of a legacy micro-service-based application [35]. In other words, we devised a method to classify services composing the application as *edgifiable* or not, based on several criteria. We applied this method to the particular case of the ShareLatex application which enables the collaborative edition of LaTeX documents. Recently, we continue this work by automate the localization and the migration of microservices. Our middleware, based on Koala [36], a lightweight Distributed Hash Table, allows adapting compatible legacy microservices applications for hybrid core/edge deployments [21].

7.1.4. Community Clouds

Participants: Jean-Louis Pazat, Bruno Stevant.

Small communities of people who need to share data and applications can now buy inexpensive devices in order to use only "on premise" resources instead of public Clouds. This "self-hosting-and-sharing" solution provides a better privacy and does not need people to pay any monthly fee to a resource provider. We have implemented a prototype based on micro-services in order to be able to distribute the load of applications among devices.

However, such a distributed platform needs to rely on a very good distribution of the computing and communication load over the devices. Using an emulator of the system, we have shown that, thanks to well known optimization techniques (Particle Swarm Optimization), it is possible to quickly find a service placement resulting in a response time close to the optimal one.

This year we evaluated the results of the optimization algorithm on a prototype (5 "boxes" installed in different home locations connected by fiber or ADSL). Results shown that due to the variation of the network available bandwidth it is necessary to dynamically modify the deployment of applications. This was not a big surprise, but we were not able to find any predictive model of this variation during a day. So, we developed and experimented a dynamic adaptation of the placement of micro-services based applications based on a regular monitoring of the response time of applications. We plan to submit a paper on this topic in early 2020.

7.1.5. Geo-distributed data stream processing

Participants: Hamidreza Arkian, Davaadorj Battulga, Mehdi Belkhiria, Guillaume Pierre, Cédric Tedeschi.

We investigated a decentralized scaling mechanism for stream processing applications where the different operators composing the processing topology are able to take their own scaling decisions independently, based on local information. We built a simulation tool to validate the ability of our algorithm to react to load variation. Then, we started the development of a software prototype of a decentralized Stream Processing Engine including this autoscaling mechanism, and deployed it over the Grid'5000 platform. Two papers have been accepted in 2019 about this work [11], [12].

Although data stream processing platforms such as Apache Flink are widely recognized as an interesting paradigm to process IoT data in fog computing platforms, the existing performance model to capture of stream processing in geo-distributed environments are theoretical works only, and have not been validated against empirical measurements. We developed and experimentally validated such a model to represent the performance of a single stream processing operator [10]. This model is very accurate with predictions $\pm 2\%$ of the actual values even in the presence of heterogeneous network latencies. Individual operator models can be composed together and, after the initial calibration of a first operator, a reasonably accurate model for other operators can be derived from a single measurement only.

7.1.6. QoS-aware and energy-efficient resource management for Function-as-a-Service

Participants: Yasmina Bouizem, Christine Morin, Nikos Parlavantzas.

Recent years have seen the widespread adoption of serverless computing, and in particular, Function-as-a-Service (FaaS) systems. These systems enable users to execute arbitrary functions without managing underlying servers. However, existing FaaS frameworks provide no quality of service guarantees to FaaS users in terms of performance and availability. Moreover, they provide no support for FaaS providers to reduce energy consumption. The goal of this work is to develop an automated resource management solution for FaaS plaforms that takes into account performance, availability, and energy efficiency in a coordinated manner. This work is performed in the context of the thesis of Yasmina Bouizem. In 2019, we integrated a fault-tolerance mechanism into Fission, an open-source FaaS framework based on Kubernetes, and are currently evaluating its impact on performance, availability, and energy consumption.

7.2. Greening Clouds

7.2.1. Energy Models

Participants: Loic Guegan, Anne-Cécile Orgerie, Martin Quinson.

Cloud computing allows users to outsource the computer resources required for their applications instead of using a local installation. It offers on-demand access to the resources through the Internet with a pay-as-you-go pricing model. However, this model hides the electricity cost of running these infrastructures.

The costs of current data centers are mostly driven by their energy consumption (specifically by the air conditioning, computing and networking infrastructures). Yet, current pricing models are usually static and rarely consider the facilities' energy consumption per user. The challenge is to provide a fair and predictable model to attribute the overall energy costs per virtual machine and to increase energy-awareness of users. We aim at proposing such energy cost models without heavily relying on physical wattmeters that may be costly to install and operate. These results have been published in [24].

Another goal consists in better understanding the energy consumption of computing and networking resources of Clouds in order to provide energy cost models for the entire infrastructure including incentivizing cost models for both Cloud providers and energy suppliers. These models should be based on experimental measurement campaigns on heterogeneous devices. As hardware architectures become more complex, measurement campains are required to better understand their energy consumption and to identify potential sources of energy waste. These results, conducted with Amina Guermouche (IMT Telecom SudParis), have been presented in [30].

Similarly, software stacks add complexity in the identification of energy inefficiencies. For HPC applications, precise measurements are required to determine the most efficient options for the runtime, the resolution algorithm and the mapping on physical resources. An example of such a study has been published in collaboration with HiePACS (Bordeaux) and NACHOS (Sophia) teams in [8].

The fine-grain measurements lead us to propose models that have been used to compare different Cloud architectures (from fog and edge to centralized clouds) in terms of energy consumption on a given scenario. These results have been published in [4].

Inferring a cost model from energy measurements is an arduous task since simple models are not convincing, as shown in our previous work. We aim at proposing and validating energy cost models for the heterogeneous Cloud infrastructures in one hand, and the energy distribution grid on the other hand. These models will be integrated into simulation frameworks in order to validate our energy-efficient algorithms at larger scale. In particular, this year we implemented in SimGrid a flow-based energy model for wired network devices [17].

7.2.2. End-to-end energy models for the Internet of Things

Participants: Anne-Cécile Orgerie, Loic Guegan.

The development of IoT (Internet of Things) equipment, the popularization of mobile devices, and emerging wearable devices bring new opportunities for context-aware applications in cloud computing environments. The disruptive potential impact of IoT relies on its pervasiveness: it should constitute an integrated heterogeneous system connecting an unprecedented number of physical objects to the Internet. Among the many challenges raised by IoT, one is currently getting particular attention: making computing resources easily accessible from the connected objects to process the huge amount of data streaming out of them.

While computation offloading to edge cloud infrastructures can be beneficial from a Quality of Service (QoS) point of view, from an energy perspective, it is relying on less energy-efficient resources than centralized Cloud data centers. On the other hand, with the increasing number of applications moving on to the cloud, it may become untenable to meet the increasing energy demand which is already reaching worrying levels. Edge nodes could help to alleviate slightly this energy consumption as they could offload data centers from their overwhelming power load and reduce data movement and network traffic. In particular, as edge cloud infrastructures are smaller in size than centralized data center, they can make a better use of renewable energy.

We investigate the end-to-end energy consumption of IoT platforms. Our aim is to evaluate, on concrete use-cases, the benefits of edge computing platforms for IoT regarding energy consumption. We aim at proposing end-to-end energy models for estimating the consumption when offloading computation from the objects to the Cloud, depending on the number of devices and the desired application QoS. This work has been published in [18].

7.2.3. Exploiting renewable energy in distributed clouds

Participants: Benjamin Camus, Anne-Cécile Orgerie.

The growing appetite of Internet services for Cloud resources leads to a consequent increase in data center (DC) facilities worldwide. This increase directly impacts the electricity bill of Cloud providers. Indeed, electricity is currently the largest part of the operation cost of a DC. Resource over-provisioning, energy non-proportional behavior of today's servers, and inefficient cooling systems have been identified as major contributors to the high energy consumption in DCs.

In a distributed Cloud environment, on-site renewable energy production and geographical energy-aware load balancing of virtual machines allocation can be associated to lower the brown (i.e. not renewable) energy consumption of DCs. Yet, combining these two approaches remains challenging in current distributed Clouds. Indeed, the variable and/or intermittent behavior of most renewable sources – like solar power for instance – is not correlated with the Cloud energy consumption, that depends on physical infrastructure characteristics and fluctuating unpredictable workloads.

7.2.4. Smart Grids

Participants: Anne Blavette, Benjamin Camus, Anne-Cécile Orgerie, Martin Quinson.

Smart grids allow to efficiently perform demand-side management in electrical grids in order to increase the integration of fluctuating and/or intermittent renewable energy sources in the energy mix. In this work, we consider the computing infrastructure that controls the smart grid. This infrastructure comprises communication and computing resources to allow for a smart management of the electrical grid. In particular, we study the influence of communication latency over a shedding scenario on a small-scale electrical network. We show that depending on the latency some shedding strategies are not feasible [13].

7.3. Securing Clouds

7.3.1. Security monitoring in Cloud computing platforms

Participants: Clément Elbaz, Christine Morin, Louis Rilling, Amir Teshome Wonjiga.

In the INDIC project we aim at making security monitoring a dependable service for IaaS cloud customers. To this end, we study three topics:

- defining relevant SLA terms for security monitoring,
- enforcing and verifying SLA terms,
- making the SLA terms enforcement mechanisms self-adaptable to cope with the dynamic nature of clouds.

The considered enforcement and verification mechanisms should have a minimal impact on performance.

In the past years we proposed a verification method for security monitoring SLOs [37] and we have then studied a methodology to define security monitoring SLOs that are at the same time relevant for the tenant, achievable for the provider, and verifiable. The methodology is based on metrics benchmarks that a cloud service provider runs on a set of basic setups of an NIDS (Network Intrusion Detection), the basic setups covering together the variety of NIDS rules that may interest tenants. In order to make it achievable for a cloud service provider to run such benchmarks despite thousands of rules that could be chosen individually by tenants, we proposed a rule clustering strategy to lower the number of sets of rules that should be benchmarked and thus the number of benchmarks run. Finally we proposed extensions to an existing cloud SLA language to define security monitoring SLOs. These results were published in a technical report [27] as well as in Amir Teshome Wonjiga's thesis (to appear) and were submitted for publication in an international conference.

In a side project with Dr Sean Peisert at LBNL, the work on security SLO verification was extended to the use case of data integrity, where tenants outsource data to a cloud storage provider. This work allowed us to tackle a challenge in SLO verification because, in this use case as well as in the security monitoring use case, tenants cannot verify SLOs without a minimal trust in providers involvment in the verification process. We proposed a strategy based on blockchains that allows tenants as well as providers to do SLO verification without having to trust any individual entity. This work was published in the CIFS security workshop [22].

To make security monitoring SLOs adaptable to context changes like the evolution of threats and updates to the tenants' software, we have worked on automating the mitigation of new threats during the time window in which no intrusion detection rule exist and no security patch is applied yet (if available). This time winwow is critical because newly published vulnerabilities get exploited up to five orders of magnitude right after they are published and the time window may last several days or weeks. We have worked on a first step of mitigation, which consists in deciding if a newly published vulnerabiliy impacts a given information system. A major challenge in automating this step is that newly published vulnerabilities do not contain machine-readable data and this data only appears up to several weeks later. For this reason we designed and evaluated a keyword extraction process from the free-form text description of a vulnerability to map a given vulnerability to product names. This keyword exctraction process was first published at the RESSI French security conference [23] and will appear in the NOMS 2020 international conference. In future work this mapping should be combined with a knowledge base of the information system to automatically score the impact of a new vulnerability on the information system.

Our results were published in [27], [28], [22], [23], [25].

7.3.2. Privacy monitoring in Fog computing platforms

Participants: Mozhdeh Farhadi, Guillaume Pierre.

IoT devices are integrated in our daily lives, and as a result they often have access to lots of private information. For example many digital assistants (Alexa, Amazon Echo...) were shown to have violated the privacy policy they had established themselves. To increase the level of confidence that end users may have in these devices and the applications which process their data, we started designing monitoring mechanisms such that the fog or the cloud platform can certify whether an application actually follows its own privacy policy or not. A survey paper on security of fog computing platforms is under submission, and we expect another paper on privacy monitoring in 2020.

7.4. Experimenting with Clouds

7.4.1. Simulating distributed IT systems

Participants: Toufik Boubehziz, Benjamin Camus, Anne-Cécile Orgerie, Millian Poquet, Martin Quinson.

Our team plays a major role in the advance of the SimGrid simulator of IT systems. This framework has a major impact on the community. Cited by over 900 papers, it was used as a scientific instrument by more than 300 publications over the years.

This year, we pursued our effort to ensure that SimGrid becomes a *de facto* standard for the simulation of distributed IT platforms. We further polished the new interface to ensure that it correctly captures the concepts needed by the experimenters, and provided a Python binding to smooth the learning curve. To that extend, we also continued our rewriting of the documentation.

The work on SimGrid is fully integrated to the other research efforts of the Myriads team. This year, we added the ability to co-simulate IT systems with SimGrid and physical systems modeled with equational systems [13]. This work, developed to study the co-evolution of thermal systems or of the electic grid with the IT system, is now distributed as an official plugin of the SimGrid framework.

7.4.2. Formal methods for IT systems

Participants: Ehsan Azimi, The Anh Pham, Martin Quinson.

The SimGrid framework also provide a state of the art Model-Checker for MPI applications. This can be used to formally verify whether the application entails synchronization issues such as deadlocks or livelocks [32]. This year, we pursued our effort on this topic, in collaboration with Thierry Jéron (EPI SUMO).

The Anh Pham defended his thesis this year on techniques to mitigate the state space explosion while verifying asynchronous distributed applications. He adapted an algorithm leveraging event folding structures to this context. This allows to efficiently compute how to not explore equivalent execution traces more than once. This work was published this year[19]. This work, co-advised by Martin Quinson with Thierry Jéron (team SUMO, formal methods), was important to bridge the gap between the involved communities.

Ehsan Azimi joined the Myriads team as an engineer in December to integrate the results of this thesis into the SimGrid framework.

7.4.3. Executing epidemic simulation applications in the Cloud

Participants: Christine Morin, Nikos Parlavantzas, Manh Linh Pham.

In the context of the DiFFuSE ADT and in collaboration with INRA researchers, we transformed a legacy application for simulating the spread of Mycobacterium avium subsp. paratuberculosis (MAP) to a cloud-enabled application based on the DiFFuSE framework (Distributed framework for cloud-based epidemic simulations). This is the second application to which the DiFFuSE framework is applied. The first application was a simulator of the spread of the bovine viral diarrhea virus, developed within the MIHMES project (2012-2017). Using both the MAP and BVDV applications, we performed extensive experiments showing the advantages of the DiFFuSE framework. Specifically, we showed that DiFFuSE enhances application performance and allows exploring different cost-performance trade-offs while supporting automatic failure handling and elastic resource acquisition from multiple clouds [7].

7.4.4. Tools for experimentation

Participant: Matthieu Simonin.

In collaboration with the STACK team and in the context of the Discovery IPL, novel experimentation tools have been developed. In this context experimenting with large software stacks (OpenStack, Kubernetes) was required. These stacks are often tedious to handle. However, practitioners need a right abstraction level to express the moving nature of experimental targets. This includes being able to easily change the experimental conditions (e.g underlying hardware and network) but also the software configuration of the targeted system (e.g service placement, fined-grained configuration tuning) and the scale of the experiment (e.g migrate the experiment from one small testbed to another bigger testbed).

In this spirit we discuss in [31] a possible solution to the above desiderata. We illustrate its use in a real world use case study which has been completed in [34]. We show that an experimenter can express their experimental workflow and execute it in a safe manner (side effects are controlled) which increases the repeatability of the experiments.

The outcome is a library (EnOSlib) target reusability in experiment driven research in distributed systems. The library can be found in https://bil.inria.fr/fr/software/view/3589/tab.

STACK Project-Team

7. New Results

7.1. Resource Management

Participants: Mohamed Abderrahim, Adwait Jitendra Bauskar, Emile Cadorel, Hélène Coullon, Jad Darrous, David Espinel, Shadi Ibrahim, Thomas Lambert, Adrien Lebre, Jean-Marc Menaud, Alexandre Van Kempen.

In 2019, we achieved several contributions regarding the management of resources and data of cloud infrastructures, especially in a geo-distributed context (*e.g.*, Fog and Edge computing).

The first contributions are related to improvements of low-level building blocks. The following ones deal with geo-distributed considerations. Finally the last ones are related to capacity and placement strategies of distributed applications and scientific workflows.

In [15], we discuss how to improve I/O fairness and SSDs' utilization through the introduction of a NCQ-aware I/O scheduling scheme, NASS. The basic idea of NASS is to elaborately control the request dispatch of workloads to relieve NCQ conflict and improve NCQ utilization at the same time. To do so, NASS builds an evaluation model to quantify important features of the workload. In particular, the model first finds aggressive workloads, which cause NCQ conflict, based on the request size and the number of requests of the workloads. Second, it evaluates merging tendency of each workload, which may affect the bandwidth and cause NCQ conflict indirectly, based on request merging history. Third, the model identifies workloads with deceptive idleness, which cause low NCQ utilization, based on historical requests in I/O scheduler. Then, based on the model, NASS sets the request dispatch of each workload to guarantee fairness and improve device utilization: (1) NASS limits aggressive workloads to relieve NCQ conflict; (2) it adjusts merging of sequential workloads to improve bandwidth of the workloads while relieving NCQ conflict; and (3) it restricts request dispatch of I/O scheduler, rather than stopping request dispatch to improve NCQ utilization. We integrate NASS into four state-of-the-art I/O schedulers including CFQ, BFQ, FlashFQ, and FIOPS. The experimental results show that with NASS, I/O schedulers can achieve 11-23% better fairness and at the same time improve device utilization by 9-29%.

In [16], [28], we address the challenge related to the boot duration of virtual machines and containers in high consolidated cloud scenarios.This time, which can last up to minutes, is critical as it defines how an application can react w.r.t. demands' fluctuations (horizontal elasticity). Our contribution is the YOLO proposal (You Only Load Once). YOLO reduces the number of I/O operations generated during a boot process by relying on a boot image abstraction, a subset of the VM/container image that contains data blocks necessary to complete the boot operation. Whenever a VM or a container is booted, YOLO intercepts all read accesses and serves them directly from the boot image, which has been locally stored on fast access storage devices (e.g., memory, SSD, etc.). In addition to YOLO, we show that another mechanism is required to ensure that files related to VM/container management systems remain in the cache of the host OS. Our results show that the use of these two techniques can speed up the boot duration 2–13 times for VMs and 2 times for containers. The benefit on containers is limited due to internal choices of the docker design. We underline that our proposal can be easily applied to other types of virtualization (e.g., Xen) and containerization because it does not require intrusive modifications on the virtualization/container management system nor the base image structure.

Complementary to the previous contribution and in an attempt to demonstrate the importance of container image placement across edge servers, we propose and evaluate through simulation two novel container image placement algorithms based on k-Center optimization in [14]. In particular, we introduce a formal model to tackle down the problem of reducing the maximum retrieval time of container images, which we denote as MaxImageRetrievalTime. Based on the model, we propose KCBP and KCBP-WC, two placement algorithms which target reducing the maximum retrieval time of container images from any edge server. While KCBP is based on a k-Center solver (i.e., placing k facilities on a set of nodes to minimize the distance from any node to the closet facility) which is applied on each layer and its replicas (taking into account the storage capacities

of the nodes), KCBP-WC uses the same principle but it tries to avoid simultaneous downloads from the same node. More precisely, if two layers are part of the same image, then they cannot be placed on the same nodes. We have implemented our proposed algorithms alongside two other state-of-the-art placement algorithms (i.e., Best-Fit and Random) in a simulator written in Python. Simulation results show that the proposed algorithms can outperform state-of- the-art algorithms by a factor of 1.1x to 4x depending on the characteristics of the networks.

In [13], we conduct experiments to thoroughly understand the performance of data-intensive applications under replication and EC. We use representative benchmarks on the Grid'5000 testbed to evaluate how analytic workloads, data persistency, failures, the back-end storage devices, and the network configuration impact their performances. While some of our results follow our intuition, others were unexpected. For example, disk and network contentions caused by chunks distribution and the unawareness of their functionalities are the main factor affecting the performance of data-intensive applications under EC, not data locality. An important outcome of our study is that it illustrates in practice the potential benefits of using EC in data-intensive clusters, not only in reducing the storage cost – which is becoming more critical with the wide adoption of high-speed storage devices and the explosion of generated and to be processed data – but also in improving the performance of data-intensive applications. We extended our work to Fog infrastructures in [31]. In particular, we empirically demonstrate the impact of network heterogeneity on the execution time of MR applications when running in the Fog.

In [5], we propose a first approach to deal with the data location challenges in geo-distribtued object stores. Existing solutions, relying on a distributed hash table to locate the data, are not efficient because location record may be placed far away from the object replicas. In this work, we propose to use a tree-based approach to locate the data, inspired by the Domain Name System (DNS) protocol. In our protocol, servers look for the location of an object by requesting successively their ancestors in a tree built with a modified version of the Dijkstra's algorithm applied to the physical topology. Location records are replicated close to the object replicas to limit the network traffic when requesting an object. We evaluate our approach on the Grid'5000 testbed using micro experiments with simple network topologies and a macro experiment using the topology of the French National Research and Education Network (RENATER). In this macro benchmark, we show that the time to locate an object in our approach is less than 15 ms on average which is around 20% shorter than using a traditional Distributed Hash Table (DHT).

In [20], we present the design, implementation, and evaluation of F-Storm, an FPGA-accelerated and general-purpose distributed stream processing system in the Edge. By analyzing current efforts to enable stream data processing in the Edge and to exploit FPGAs for data-intensive applications, we derive the key design aspects of F-Storm. Specifically, F-Storm is designed to: (1) provide a light-weight integration of FPGA with a DSP system in Edge servers, (2) make full use of FPGA resources when assigning tasks, (3) relieve the high overhead when transferring data between Java Virtual Machine (JVM) and FPGAs, and importantly (4) provide programming interface for users that enable them to leverage FPGA accelerators easily while developing their stream data applications. We have implemented F-Storm based on Storm. Evaluation results show that F-Storm reduces the latency by 36% and 75% for matrix multiplication and grep application compared to Storm. Furthermore, F-Storm obtains 1.4x, 2.1x, and 3.4x throughput improvement for matrix multiplication, grep application, and vector addition, respectively.

In [30], we discuss the main challengres related to the design and development of inter-site services for operating a massively distributed Cloud-Edge architecture deployed in different locations of the Internet backbone (i.e, network point of presences). More precisely, we discuss challenges related to the establishment of connectivity among several virtual infrastructure managers in charge of operating each site. Our goal is to initiate the discussion about the research directions on this field providing some interesting points to promote future work.

In [7], we focus on how to reduce the costly cross-rack data transferring in MapReduce systems. We observe that with high Map locality, the network is mainly saturated in Shuffling but relatively free in the Map phase. A little sacrifice in Map locality may greatly accelerate Shuffling. Based on this, we propose a novel scheme called Shadow for Shuffle-constrained general applications, which strikes a trade-off between Map locality and

Shuffling load balance. Specifically, Shadow iteratively chooses an original Map task from the most heavily loaded rack and creates a duplicated task for it on the most lightly loaded rack. During processing, Shadow makes a choice between an original task and its replica by efficiently pre-estimating the job execution time. We conduct extensive experiments to evaluate the Shadow design. Results show that Shadow greatly reduces the cross-rack skewness by 36.6% and the job execution time by 26% compared to existing schemes.

In [6], we consider a complete framework for straggler detection and mitigation. We start with a set of metrics that can be used to characterize and detect stragglers including Precision, Recall, Detection Latency, Undetected Time and Fake Positive. We then develop an architectural model by which these metrics can be linked to measures of performance including execution time and system energy overheads. We further conduct a series of experiments to demonstrate which metrics and approaches are more effective in detecting stragglers and are also predictive of effectiveness in terms of performance and energy efficiencies. For example, our results indicate that the default Hadoop straggler detector could be made more effective. In certain case, Precision is low and only 55% of those detected are actual stragglers and the Recall, i.e., percent of actual detected stragglers, is also relatively low at 56%. For the same case, the hierarchical approach (i.e., a green-driven detector based on the default one) achieves a Precision of 99% and a Recall of 29%. This increase in Precision can be translated to achieve lower execution time and energy consumption, and thus higher performance and energy efficiency; compared to the default Hadoop mechanism, the energy consumption is reduced by almost 31%. These results demonstrate how our framework can offer useful insights and be applied in practical settings to characterize and design new straggler detection mechanisms for MapReduce systems.

In [21], we provide a general solution for workflow performance optimizations considering system variations. Specifically, we model system variations as time-dependent random variables and take their probability distributions as optimization input. Despite its effectiveness, this solution involves heavy computation overhead. Thus, we propose three pruning techniques to simplify workflow structure and reduce the probability evaluation overhead. We implement our techniques in a runtime library, which allows users to incorporate efficient probabilistic optimization into existing resource provisioning methods. Experiments show that probabilistic solutions can improve the performance by 51% compared to state-of-the-art static solutions while guaranteeing budget constraint, and our pruning techniques can greatly reduce the overhead of probabilistic optimization.

In [11], we propose a new strategy to schedule heteregeneous scientific workflows while minimizing the energy consumption of the cloud provider by introducing a deadline sensitive algorithm. Scheduling workflows in a cloud environment is a difficult optimization problem as capacity constraints must be fulfilled additionally to dependencies constraints between tasks of the workflows. Usually, work around the scheduling of scientific workflows focuses on public clouds where infrastructure management is an unknown black box. Thus, many works offer scheduling algorithms designed to select the best set of virtual machines over time, so that the cost to the end user is minimized. This paper presents the new v-HEFT-*deadline* algorithm that takes into account users deadlines to minimize the number of machines used by the cloud provider. The results show the real benefits of using our algorithm for reducing the energy consumption of the cloud provider.

In [9],we investigate how a monitoring service for Edge infrastructures should be designed in order to mitigate as much as possible its footprint in terms of used resources. Monitoring functions tend to become compute-, storage-and network-intensive, in particular because they will be used by a large part of applications that rely on real-time data. To reduce as much as possible the footprint of the whole monitoring service, we propose to mutualize identical processing functions among different tenants while ensuring their quality-of-service (QoS) expectations. We formalize our approach as a constraint satisfaction problem and show through micro-benchmarks its relevance to mitigate compute and network footprints.

In [22], we propose a generalization of the previous work. More precisely, weinvestigates whether the use of Constraint Programming (CP) could enable the development of a generic and easy-to-upgrade placement service for Fog/Edge Computing infrastructures. Our contribution is a new formulation of the placement problem, an implementation of this model leveraging Choco-solver and an evaluation of its scalability in comparison to recent placement algorithms. To the best of our knowledge, our study is the first one to evaluate

the relevance of CP approaches in comparison to heuristic ones in this context. CP interleaves inference and systematic exploration to search for solutions, letting users on what matters: the problem description. Thus, our service placement model not only can be easily enhanced (deployment constraints/objectives) but also shows a competitive tradeoff between resolution times and solutions quality.

In [27], we present the first building blocks of a simulator to investigate placement challenges in Edge infrastructures. Efficiently scheduling computational jobs with data-sets dependencies is one of the most important challenges of fog/edge computing infrastructures. Although several strategies have been proposed, they have been evaluated through ad-hoc simulator extensions that are, when available, usually not maintained. This is a critical problem because it prevents researchers to-easily-conduct fair evaluations to compare each proposal. We propose to address this limitation throught the design and development of a common simulator. More precisely, in this research report, we describe an ongoing project involving academics and a high-tech company that aims at delivering a dedicated tool to evaluate scheduling policies in edge computing infrastructures. This tool enables the community to simulate various policies and to easily customize researchers/engineers' use-cases, adding new functionalities if needed. The implementation has been built upon the Batsim/SimGrid toolkit, which has been designed to evaluate batch scheduling strategies in various distributed infrastructures. Although the complete validation of the simulation toolkit is still ongoing, we demonstrate its relevance by studying different scheduling strategies on top of a simulated version of the Qarnot Computing platform, a production edge infrastructure based on smart heaters.

In [8], we propose an efficient graph partitioning method named Geo-Cut, which takes both the cost and performance objectives into consideration for large graph processing in geo-distributed DCs.Geo-Cut adopts two optimization stages. First, we propose a cost-aware streaming heuristic and utilize the one-pass streaming graph partitioning method to quickly assign edges to different DCs while minimizing inter-DC data communication cost. Second, we propose two partition refinement heuristics which identify the performance bottlenecks of geo-distributed graph processing and refine the partitioning result obtained in the first stage to reduce the inter-DC data transfer time while satisfying the budget constraint. Geo-Cut can be also applied to partition dynamic graphs thanks to its lightweight runtime overhead. We evaluate the effectiveness and efficiency of Geo-Cut using real-world graphs with both real geo-distributed DCs and simulations. Evaluation results show that Geo-Cut can reduce the inter-DC data transfer time by up to 79% (42% as the median) and reduce the monetary cost by up to 75% (26% as the median) compared to state-of-the-art graph partitioning methods with a low overhead.

7.2. Programming Support

Participants: Maverick Chardet, Hélène Coullon, Thomas Ledoux, Jacques Noyé, Dimitri Pertin, Simon Robillard, Hamza Sahli, Charlène Servantie.

Our contributions regarding programming support are divided in two topics. First, we focused on one specific challenge related to distributed software deployment: distributed software commissioning. We have proposed a useful approach for introducing model checking to help system operators design their parallel distributed software commissioning. Then, we focused on Fog formalization and we have proposed a fully graphical process algebraic formalism to design a Fog system.

In [12], MADA, a deployment approach to facilitate the design of efficient and safe distributed software commissioning is presented. MADA is built on top of the Madeus formal model that focuses on the efficient execution of installation procedures. Madeus puts forward more parallelism than other commissioning models, which implies a greater complexity and a greater propensity for errors. MADA provides a new specific language on top of Madeus that allows the developer to easily define the properties that should be ensured during the commissioning process. Then, MADA automatically translates the description to a time Petri net and a set of TCTL formulae. MADA is evaluated on the OpenStack commissioning.

About Fog formalization, we present a novel formal model defining spatial and structural aspects of Fog-based systems using Bigraphical Reactive Systems, a fully graphical process algebraic formalism [17]. The model is extended with reaction rules to represent the dynamic behavior of Fog systems in terms of self-adaptation.

The notion of bigraph patterns is used in conjunction with boolean and temporal operators to encode spatio-temporal properties inherent to Fog systems and applications. The feasibility of the modelling approach is demonstrated via a motivating case study and various self-adaptation scenarios.

Overall, the number of contributions we made this year on the programming support topic is less significative than the previous one. However, we would like to underline that it does not reflect the recent efforts we put. In particular, the team has strongly developed the field of dynamic reconfiguration of distributed software systems and expects to get important results during 2020.

7.3. Energy-aware computing

Participants: Emile Cadorel, Hélène Coullon, Adrien Lebre, Thomas Ledoux, Jean-Marc Menaud, Jonathan Pastor, Dimitri Saingre, Yewan Wang.

Energy consumption is one of the major challenges of modern datacenters and supercomputers. Our works in Energy-aware computing can be categorized into two subdomains: Software level (SaaS, PaaS) and Infrastructure level (IaaS). At Software level, we worked on the general Cloud applications architectures and more recently on BlockChain-based solutions. At Infrastructure level, we worked this year on two directions: (i) investigating the thermal aspects in datacenters, and (ii) analyzing the energy footprint of geo-distributed plateforms.

In [11], the scheduling of heterogeneous scientific workflows while minimizing the energy consumption of the cloud provider is tackled by introducing a deadline sensitive algorithm. Scheduling workflows in a cloud environment is a difficult optimization problem as capacity constraints must be fulfilled additionally to dependencies constraints between tasks of the workflows. Usually, work around the scheduling of scientific workflows focuses on public clouds where infrastructure management is an unknown black box. Thus, many works offer scheduling algorithms designed to select the best set of virtual machines over time, so that the cost to the end user is minimized. This paper presents the new v-HEFT-*deadline* algorithm that takes into account users deadlines to minimize the number of machines used by the cloud provider. The results show the real benefits of using our algorithm for reducing the energy consumption of the cloud provider.

In [25], over the last year, both academic and industry have increase their work on blockchain technologies. Despite the potential of blockchain technologies in many areas, several obstacles are slowing down their development. In addition to the legal and social obstacles, technical limitations now prevent them from imposing themselves as a real alternative to centralised services. For example, several problems dealing with the scalability or the energy cost have been identified. That's why, a significant part of this research is focused on improving the performances (latency, throughput, energy footprint, etc.) of such systems. Unfortunately, Those projects are often evaluated with ad hoc tools and experimental environment, preventing reproducibility and easy comparison of new contribution to the state of the art. As a result, we notice a clear lack of tooling concerning the benchmarking of blockchain technologies. To the best of our knowledge only a few tools address such issues. Those tools often relies on the load generation aspect and omit some other important aspect of benchmark experiments such as reproducibility and the network emulation. We introduce BCTMark, a general framework for benchmarking blockchain technologies in an emulated environment in a reproductible way.

In [18], we present a deep evaluation about the power models based on CPU utilization. The influence of inlet temperature on models has been especially discussed. According to the analysis, one regression formula by using CPU utilization as the only indicator is not adequate for building reliable power models. First of all, Workloads have different behaviors by using CPU and other hardware resources in server platforms. Therefore, power is observed to have high dispersion for a fixed CPU utilization, especially at full workload. At the same time, we also find that, power is well proportional to CPU utilization within the execution of one single workload. Hence, applying workload classifications could be an effective way to improve model accuracy. Moreover, inlet temperature can cause surprising influence on model accuracy. The model reliability can be questioned without including inlet temperature data. In a use case, after including inlet temperature data, we have greatly improved the precision of model outputs while stressing server under three different ambient temperatures.

In [18], our physical experiments have shown that even under the same conditions, identical processors consume different amount of energy to complete the same task. While this manufacturing variability has been observed and studied before, there is lack of evidence supporting the hypotheses due to limited sampling data, especially from the thermal characteristics. In this article, we compare the power consumption among identical processors for two Intel processors series with the same TDP (Thermal Design Power) but from different generations. The observed power variation of the processors in newer generation is much greater than the older one. Then, we propose our hypotheses for the underlying causes and validate them under precisely controlled environmental conditions. The experimental results show that, with the increase of transistor densities, difference of thermal characteristics becomes larger among processors, which has non-negligible contribution to the variation of power consumption for modern processors. This observation reminds us of re-calibrating the precision of the current energy predictive models. The manufacturing variability has to be considered when building energy predictive models for homogeneous clusters.

In [3], we propose a model and a first implementation of a simulator in order to compare the energy footprint of different cloud architectures (single sites vs fully decentrlaized). Despite the growing popularity of Fog/Edge architectures, their energy consumption has not been well investigated yet. To move forward on such a critical question, we first introduce a taxonomy of different Cloud-related architectures. From this taxonomy, we then present an energy model to evaluate their consumption. Unlike previous proposals, our model comprises the full energy consumption of the computing facilities, including cooling systems, and the energy consumption of network devices linking end users to Cloud resources. Finally, we instantiate our model on different Cloud-related architectures, ranging from fully centralized to completely distributed ones, and compare their energy consumption. The results validates that a completely distributed architecture, because of not using intra-data center network and large-size cooling systems, consumes less energy than fully centralized and partly distributed architectures respectively. To the best of our knowledge, our work is the first one to propose a model that enables researchers to analyze and compare energy consumption of different Cloud-related architectures.

7.4. Security and Privacy

Participants: Mohammad-Mahdi Bazm, Fatima Zahra Boujdad, Wilmer Edicson Garzon Alfonso, Jean-Marc Menaud, Sirine Sayadi, Mario Südholt.

This year the team has provided two major contributions on security and privacy challenges in distributed systems. First, we have extended our model for secure and privacy-aware biomedical analyses, as well as started to explore the impact of the big-data analyses in this context. Second, we have contributed mitigation methods for Cloud-based side-channel attacks.

In [24], we have developed a methodology for the development of secure and privacy-aware biomedical analyses we motivate the need for real distributed biomedical analyses in the context of several ongoing projects, including the I-CAN project that involves 34 French hospitals and affiliated research groups. We present a set of distributed architectures for such analyses that we have derived from discussions with different medical research groups and a study of related work. These architectures allow for scalability, security/privacy and reproducibility properties to be taken into account. A predefined set of architectures allows medecins and biomedical engineers to define high-level distributed architectures for biomedical analyses that ensure strong security and constraints on private data. Architectures from this set can then be implemented with ease because of detailed, also predefined, detailed implementation templates. Finally, we illustrate how these architectures can serve as the basis of a development method for biomedical distributed analyses.

In [10] and [23], we presented a new taxonomy for container security with a particular focus on data transmitted through the virtualization boundary. Containerization is a lightweight virtualization technique reducing virtualization overhead and deployment latency compared to full VM; its popularity is quickly increasing. However, due to kernel sharing, containers provide less isolation than full VM. Thus, a compromised container may break out of its isolated context and gain root access to the host server. This is a huge concern, especially in multi-tenant cloud environments where we can find running on a single server containers serving very different purposes, such as banking microservices, compute nodes or honeypots. Thus, containers with specific security needs should be able to tune their own security level. Because OS-level defense approaches

inherited from time-sharing OS generally requires administrator rights and aim to protect the entire system, they are not fully suitable to protect usermode containers. Research recently made several contributions to deliver enhanced security to containers from host OS level to (partially) solve these challenges. In this survey, we propose a new taxonomy on container defense at the infrastructure level with a particular focus on the virtualization boundary, where interactions between kernel and containers take place. We then classify the most promising defense frameworks into these categories.

Finally, we have leveraged an approach based on Moving Target Defense (MTD) theory to interrupt a cache-based side-channel attack between two Linux containers in the context of the Mohammad Mahdi's PhD thesis [1]. MTD allows us to make the configuration of system more dynamic and consequently more harder to attack by an adversary, by using shuffling at different level of system and cloud. Our approach does not need to carrying modification neither into the guest OS or the hypervisor. Experimental results show that our approach imposes very low performance overhead. We have also provided a survey on the isolation challenge and on the cache-based side-channel attacks in cloud computing infrastructures. We have developed different approaches to detect/mitigate cross-VM/cross-containers cache-based side-channel attacks. Regarding the detection of cache-based side-channel attacks, we have enabled their detection by leveraging Hardware performance Counters (HPCs) and Intel Cache Monitoring Technology (CMT) with anomaly detection approaches to identify a malicious virtual machine or a Linux container. Our experimental results show a high detection rate.

WIDE Project-Team

6. New Results

6.1. Recommender Systems

6.1.1. A Biclustering Approach to Recommender Systems

Participants: Florestan de Moor, Davide Frey.

Recommendation systems are a core component of many e-commerce industries and online services since they ease the discovery of relevant products. Because catalogs are huge, it is impossible for an individual to manually search for an item of interest, hence the need for some automatic filtering process. Many approaches exist, from content-based ones to collaborative filtering that include neighborhood and model-based techniques. Despite these intensive research activities, numerous challenges remain to be addressed, particularly under real-time settings or regarding privacy concerns, which motivates further work in this area. We focus on techniques that rely on biclustering, which consists in simultaneously building clusters over the two dimensions of a data matrix. Although it was little considered by the recommendation system community, it is a well-known technique in other domains such as genomics. In work [42] we present the different biclustering-based approaches that were explored. We then are the first to perform an extensive experimental evaluation to compare these approaches with one another, but also with the current state-of-the-art techniques from the recommender field. Existing evaluations are often restrained to a few algorithms and consider only a limited set of metrics. We then expose a few ideas to improve existing approaches and address the current challenges in the design of highly efficient recommendation algorithms, along with some preliminary results.

This work was done in collaboration with Antonio Mucherino (University of Rennes 1).

6.1.2. Unified and Scalable Incremental Recommenders with Consumed Item Packs

Participant: Erwan Le Merrer.

Recommenders personalize the web content by typically using collaborative filtering to relate users (or items) based on explicit feedback, e.g., ratings. The difficulty of collecting this feedback has recently motivated to consider implicit feedback (e.g., item consumption along with the corresponding time). In this work [39], we introduce the notion of consumed itempack (CIP) which enables to link users (or items) based on their implicit analogous consumption behavior. Our proposal is generic, and we show that it captures three novel implicit recommenders: a user-based (CIP-U), an item-based (CIP-I),and a word embedding-based (DEEPCIP), as well as a state-of-art technique using implicit feedback (FISM). We show that our recommenders handle incremental updates incorporating freshly consumed items. We demonstrate that all three recommenders provide a recommendation quality that is competitive with state-of-the-art ones, including one incorporating both explicit and implicit feedback

This work was done in collaboration with Rachid Guerraoui (EPFL), Rhicheek Patra (Oracle) and Jean-Ronan Vigouroux (Technicolor).

6.2. Systems for the Support of Privacy

6.2.1. Robust Privacy-Preserving Gossip Averaging

Participants: Amaury Bouchra-Pilet, Davide Frey, François Taïani.

This contribution aims to address the privacy risks inherent in decentralized systems by considering the emblematic problem of privacy-preserving decentralized averaging. In particular, we propose a novel gossip protocol that exchanges noise for several rounds before starting to exchange actual data. This makes it hard for an honest but curious attacker to know whether a user is transmitting noise or actual data. Our protocol and analysis do not assume a lock-step execution, and demonstrate improved resilience to colluding attackers. In a paper, publishing this work at SSS 2019 [26], we prove the correctness of this protocol as well as several privacy results. Finally, we provide simulation results about the efficiency of our averaging protocol.

6.2.2. A Collaborative Strategy for Mitigating Tracking through Browser Fingerprinting.

Participants: David Bromberg, Davide Frey, Alejandro Gomez-Boix.

Browser fingerprinting is a technique that collects information about the browser configuration and the environment in which it is running. This information is so diverse that it can partially or totally identify users online. Over time, several countermeasures have emerged to mitigate tracking through browser fingerprinting. However, these measures do not offer full coverage in terms of privacy protection, as some of them may introduce inconsistencies or unusual behaviors, making these users stand out from the rest.

In this work, we address these limitations by proposing a novel approach that minimizes both the identifiability of users and the required changes to browser configuration. To this end, we exploit clustering algorithms to identify the devices that are prone to share the same or similar fingerprints and to provide them with a new non-unique fingerprint. We then use this fingerprint to automatically assemble and run web browsers through virtualization within a docker container. Thus all the devices in the same cluster will end up running a web browser with an indistinguishable and consistent fingerprint.

We carried out this work in collaboration with Benoit Baudry from KTH Sweden and published our results at the 2019 Moving-Target Defense Workshop [30].

6.3. Distributed Algorithms

6.3.1. One for All and All for One:Scalable Consensus in a Hybrid Communication Model

Participant: Michel Raynal.

This work [34] addresses consensus in an asynchronous model where the processes are partitioned into clusters. Inside each cluster, processes can communicate through a shared memory, which favors efficiency. Moreover, any pair of processes can also communicate through a message-passing communication system, which favors scalability. In such a "hybrid communication" context, the work presents two simple binary consensus algorithms (one based on local coins,the other one based on a common coin). These algorithms are straightforward extensions of existing message-passing randomized round-based consensus algorithms. At each round, the processes of each cluster first agree on the same value (using an underlying shared memory consensus algorithm), and then use a message-passing algorithm to converge on the same decided value. The algorithms are such that, if all except one processes of a cluster crash, the surviving process acts as if all the processes of its cluster were alive (hence the motto "one for all and all for one"). As a consequence, the hybrid communication model allows us to obtain simple, efficient, and scalable fault-tolerant consensus algorithms. As an important side effect, according to the size of each cluster, consensus can be obtained even if a majority of processes crash.

This work was done in collaboration with Jiannong Cao (Polytechnic University, Hong Kong).

6.3.2. Optimal Memory-Anonymous Symmetric Deadlock-Free Mutual Exclusion

Participant: Michel Raynal.

The notion of an anonymous shared memory (recently introduced in PODC 2017) considers that processes use different names for the same memory location. Hence, there is permanent disagreement on the location names among processes. In this context, the PODC paper presented -among other results- a symmetric deadlock-free mutual exclusion (mutex) algorithm for two processes and a necessary condition on the size m of the anonymous memory for the existence of a symmetric deadlock-free mutex algorithm in an n-process system. This work [22] answers several open problems related to symmetric deadlock-free mutual exclusion in an n-process system where the processes communicate through m registers. It first presents two algorithms. The first considers that the registers are anonymous read/write atomic registers and works for any m greater than 1 and belonging to the set M(n). It thus shows that this condition on m is both necessary and sufficient. The second algorithm considers anonymous read/modify/write atomic registers. It assumes that $m \in M(n)$. These algorithms differ in their design principles and their costs (measured as the number of registers which must contain the identity of a process to allow it to enter the critical section). The work also shows that the condition

$m \in M(n)$ is necessary for deadlock-free mutex on top of anonymous read/modify/write atomic registers. It follows that, when m > 1, $m \in M(n)$ is a tight characterization of the size of the anonymous shared memory needed to solve deadlock-free mutex, be the anonymous registers read/write or read/modify/write.

This work was done in collaboration with Zahra Aghazadeh (University of Calgary), Damien Imbs (LIS, Université d'Aix-Marseille,CNRS, Université de Toulon), Gadi Taubenfeld (The Interdisciplinary Center of Herzliya) and Philipp Woelfel (University of Calgary).

6.3.3. Merkle Search Trees

Participants: Alex Auvolat, François Taïani.

Most recent CRDT (Conflict-free Replicated Data Type) techniques rely on a causal broadcast primitive to provide guarantees on the delivery of operation deltas. Such a primitive is unfortunately hard to implement efficiently in large open networks, whose membership is often difficult to track. As an alternative, we argue that pure state-based CRDTs can be efficiently implemented by encoding states as specialized Merkle trees, and that this approach is well suited to open networks where many nodes may join and leave. Indeed, Merkle trees enable efficient remote comparison and reconciliation of data sets, which can be used to implement the CRDT merge operator between two nodes without any prior information. This approach also does not require vector clock information, which would grow linearly with the number of participants.

At the core of our contribution [24] lies a new kind of Merkle tree, called Merkle Search Tree (MST), that implements a balanced search tree while maintaining key ordering. This latter property makes it particularly efficient in the case of updates on sets of sequential keys, a common occurrence in many applications. We use this new data structure to implement a distributed event store, and show its efficiency in very large systems with low rates of updates. In particular, we show that in some scenarios our approach is able to achieve both a 66% reduction of bandwidth cost over a vector-clock approach, as well as a 34% improvement in consistency level. We finally suggest other uses of our construction for distributed databases in open networks.

6.3.4. Dietcoin: Hardening Bitcoin Transaction Verification Process For Mobile Devices

Participants: Davide Frey, François Taïani.

Distributed ledgers are among the most replicated data repositories in the world. They offer data consistency, immutability, and auditability, based on the assumption that each participating node locally verifies their entire content. Although their content, currently extending up to a few hundred gigabytes, can be accommodated by dedicated commodity hard disks, downloading it, processing it, and storing it in general-purpose desktop and laptop computers can prove largely impractical. Even worse, this becomes a prohibitive restriction for smartphones, mobile devices, and resource-constrained IoT devices.

We thus proposed Dietcoin, a Bitcoin protocol extension that allows nodes to perform secure local verification of Bitcoin transactions with small bandwidth and storage requirements. We carried out an extensive evaluation of the features of Dietcoin that are important for today's cryptocurrency and smart-contract systems, but are missing in the current state-of-the-art. These include (i) allowing resource-constrained devices to verify the correctness of selected blocks locally without having to download the complete ledger; (ii) enabling devices to join a blockchain quickly yet securely, dropping bootstrap time from days down to a matter of seconds; (iii) providing a generic solution that can be applied to other distributed ledgers secured with Proof-of-Work. We showcased our results in a demo at VLDB 2019 [15], and we are currently preparing a full paper submission.

We carried out this work in collaboration with Pierre-Louis Roman, now at University of Lugano (Switzerland), as well as with Mark Makke from Vrije Universiteit, Amsterdam (the Netherlands), and Spyros Voulgaris from Athens University of Economics and Business (Greece).

6.3.5. Byzantine-Tolerant Set-Constrained Delivery Broadcast

Participants: Alex Auvolat, François Taïani, Michel Raynal.

Set-Constrained Delivery Broadcast (SCD-broadcast), recently introduced at ICDCN 2018, is a high-level communication abstraction that captures ordering properties not between individual messages but between sets of messages. More precisely, it allows processes to broadcast messages and deliver sets of messages, under the constraint that if a process delivers a set containing a message m before a set containing a message m', then no other process delivers first a set containing m' and later a set containing m. It has been shown that SCD-broadcast and read/write registers are computationally equivalent, and an algorithm implementing SCD-broadcast is known in the context of asynchronous message passing systems prone to crash failures.

We introduce a Byzantine-tolerant SCD-broadcast algorithm in [23], which we call BSCD-broadcast. Our proposed algorithm assumes an underlying basic Byzantine-tolerant reliable broadcast abstraction. We first introduce an intermediary communication primitive, Byzantine FIFO broadcast (BFIFO-broadcast), which we then use as a primitive in our final BSCD-broadcast algorithm. Unlike the original SCD-broadcast algorithm that is tolerant to up to $t < n/2$ crashing processes, and unlike the underlying Byzantine reliable broadcast primitive that is tolerant to up to $t < n/3$ Byzantine processes, our BSCD-broadcast algorithm is tolerant to up to $t < n/4$ Byzantine processes. As an illustration of the high abstraction power provided by the BSCD-broadcast primitive, we show that it can be used to implement a Byzantine-tolerant read/write snapshot object in an extremely simple way.

6.3.6. PnyxDB: a Lightweight Leaderless Democratic Byzantine Fault Tolerant Replicated Datastore

Participants: Loïck Bonniot, François Taïani.

Byzantine-Fault-Tolerant (BFT) systems are rapidly emerging as a viable technology for production-grade systems, notably in closed consortia deployments for financial and supply-chain applications. Unfortunately, most algorithms proposed so far to coordinate these systems suffer from substantial scalability issues, mainly due to the requirement of a single leader node. We observed that many application workloads offer little concurrency, and proposed PnyxDB, an eventually-consistent BFT replicated datastore that exhibits both high scalability and low latency. Our approach (proposed in [40]) is based on conditional endorsements, that allow nodes to specify the set of transactions that must *not* be committed for the endorsement to be valid.

Additionally, although most of prior art rely on internal voting or quorum mechanisms, these mechanisms are not exposed to applications as first-class primitives. As a result, individual nodes cannot implement application-defined policies without additional effort, costs, and complexity. This is problematic, as application-level voting capabilities are key to a number of emerging decentralized BFT applications involving independent participants who need to balance conflicting goals and shared interests. In addition to its high scalability, PnyxDB supports application-level voting by design. We provided a comparison against BFTS-MaRt and Tendermint, two competitors with different design aims, and demonstrated that our implementation speeds up commit latencies by a factor of 11, remaining below 5 seconds in a worldwide geodistributed deployment of 180 nodes.

PnyxDB's source code is freely available [0]. This work has also been done in collaboration with Christoph Neumann at InterDigital.

6.3.7. Vertex Coloring with Communication Constraints in Synchronous Broadcast Networks

Participants: Hicham Lakhlef, Michel Raynal, François Taïani.

In this work [17], we consider distributed vertex-coloring in broadcast/receive networks suffering from conflicts and collisions. (A collision occurs when, during the same round, messages are sent to the same process by too many neighbors; a conflict occurs when a process and one of its neighbors broadcast during the same round.) More specifically, our work focuses on multi-channel networks, in which a process may either broadcast a message to its neighbors or receive a message from at most γ of them. The work first provides a new upper bound on the corresponding graph coloring problem (known as frugal coloring) in general graphs, proposes an exact bound for the problem in trees, and presents a deterministic, parallel, color-optimal, collision- and conflict-free distributed coloring algorithm for trees, and proves its correctness.

[0] https://github.com/technicolor-research/pnyxdb

6.3.8. *Efficient Randomized Test-and-Set Implementations*

Participant: George Giakkoupis.

In [16], we study randomized test-and-set (TAS) implementations from registers in the asynchronous shared memory model with n processes. We introduce the problem of *group election*, a natural variant of leader election, and propose a framework for the implementation of TAS objects from group election objects. We then present two group election algorithms, each yielding an efficient TAS implementation. The first implementation has expected maxstep complexity $O(\log^* k)$ in the location-oblivious adversary model, and the second has expected maxstep complexity $O(\log \log k)$ against any read/write-oblivious adversary, where $k \leq n$ is the contention. These algorithms improve the previous upper bound by Alistarh and Aspnes (2011) of $O(\log \log n)$ expected maxstep complexity in the oblivious adversary model.

We also propose a modification to a TAS algorithm by Alistarh, Attiya, Gilbert, Giurgiu, and Guerraoui (2010) for the strong adaptive adversary, which improves its space complexity from super-linear to linear, while maintaining its $O(\log n)$ expected maxstep complexity. We then describe how this algorithm can be combined with any randomized TAS algorithm that has expected maxstep complexity $T(n)$ in a weaker adversary model, so that the resulting algorithm has $O(\log n)$ expected maxstep complexity against any strong adaptive adversary and $O(T(n))$ in the weaker adversary model.

Finally, we prove that for any randomized 2-process TAS algorithm, there exists a schedule determined by an oblivious adversary such that with probability at least $1/4^t$ one of the processes needs at least t steps to finish its TAS operation. This complements a lower bound by Attiya and Censor-Hillel (2010) on a similar problem for $n \geq 3$ processes.

This work was done in collaboration with Philipp Woelfel (University of Calgary).

6.4. Machine Learning and Security

6.4.1. *Adversarial Frontier Stitching for Remote Neural Network Watermarking*

Participant: Erwan Le Merrer.

The state-of-the-art performance of deep learning models comes at a high cost for companies and institutions, due to the tedious data collection and the heavy processing requirements. Recently, Nagai et al. proposed to watermark convolutional neural networks for image classification, by embedding information into their weights. While this is a clear progress toward model protection, this technique solely allows for extracting the watermark from a network that one accesses locally and entirely. Instead, we aim at allowing the extraction of the watermark from a neural network (or any other machine learning model) that is operated remotely, and available through a service API. To this end, we propose in this work [18] to mark the model's action itself, tweaking slightly its decision frontiers so that a set of specific queries convey the desired information. In this work, we formally introduce the problem and propose a novel zero-bit watermarking algorithm that makes use of adversarial model examples. While limiting the loss of performance of the protected model, this algorithm allows subsequent extraction of the watermark using only few queries. We experimented the approach on three neural networks designed for image classification, in the context of the MNIST digit recognition task.

This work was done in collaboration with Gilles Trédan (LAAS/CRNS) and Patrick Pérez (Valéo AI).

6.4.2. *TamperNN: Efficient Tampering Detection of Deployed Neural Nets*

Participant: Erwan Le Merrer.

Neural networks are powering the deployment of embedded devices and Internet of Things. Applications range from personal assistants to critical ones such as self-driving cars. It has been shown recently that models obtained from neural nets can be trojaned ; an attacker can then trigger an arbitrary model behavior facing crafted inputs. This has a critical impact on the security and reliability of those deployed devices. In this work [33], we introduce novel algorithms to detect the tampering with deployed models, classifiers in particular. In the remote interaction setup we consider, the proposed strategy is to identify markers of the model input space that are likely to change class if the model is attacked, allowing a user to detect a possible

tampering. This setup makes our proposal compatible with a wide range of scenarios, such as embedded models, or models exposed through prediction APIs. We experiment those tampering detection algorithms on the canonical MNIST dataset, over three different types of neural nets, and facing five different attacks (trojaning, quantization, fine-tuning, compression and watermarking). We then validate over five large models (VGG16, VGG19, ResNet, MobileNet, DenseNet) with a state of the art dataset (VGGFace2), and report results demonstrating the possibility of an efficient detection of model tampering.

This work was done in collaboration with Gilles Trédan (LAAS/CRNS).

6.4.3. MD-GAN: Multi-Discriminator Generative Adversarial Networks for Distributed Datasets

Participant: Erwan Le Merrer.

A recent technical breakthrough in the domain of machine learning is the discovery and the multiple applications of Generative Adversarial Networks (GANs). Those generative models are computationally demanding, as a GAN is composed of two deep neural networks, and because it trains on large datasets. A GAN is generally trained on a single server. In this work, we address the problem of distributing GANsso that they are able to train over datasets that are spread on multiple workers. In this work [31] MD-GAN is exposed as the first solution for this problem: we propose a novel learning procedure for GANs so that they fit this distributed setup. We then compare the performance of MD-GAN to an adapted version of Federated Learning to GANs, using the MNIST and CIFAR10 datasets.MD-GAN exhibits a reduction by a factor of two of the learning complexity on each worker node, while providing better performances than federated learning on both datasets. We finally discuss the practical implications of distributing GANs.

This work was done in collaboration with Bruno Sericola (Inria) and Corentin Hardy (Technicolor).

6.5. Network and Graph Algorithms

6.5.1. Multisource Rumor Spreading with Network Coding

Participants: David Bromberg, Quentin Dufour, Davide Frey.

The last decade has witnessed a rising interest in Gossip protocols in distributed systems. In particular, as soon as there is a need to disseminate events, they become a key functional building block due to their scalability, robustness and fault tolerance under high churn. However, Gossip protocols are known to be bandwidth intensive. A huge amount of algorithms has been studied to limit the number of exchanged messages using different combinations of push/pull approaches. In this work we revisited the state of the art by applying Random Linear Network Coding to further increase performance. In particular, the originality of our approach consists in combining sparse-vector encoding to send our network-coding coefficients and Lamport timestamps to split messages in generations in order to provide efficient gossiping. Our results demonstrate that we are able to drastically reduce bandwidth overhead and dissemination delay compared to the state of the art. We published our results at INFOCOM 2019 [27].

6.5.2. DiagNet: towards a generic, Internet-scale root cause analysis solution

Participants: Loïck Bonniot, François Taïani.

Internet content providers and network operators allocate significant resources to diagnose and troubleshoot problems encountered by end-users, such as service quality of experience degradations. Because the Internet is decentralized, the cause of such problems might lie anywhere between an end-user's device and the service datacenters. Further, the set of possible problems and causes cannot be known in advance, making it impossible to train a classifier with all combinations of faults, causes and locations. We explored how machine learning can be used for Internet-scale root cause analysis using measurements taken from end-user devices: our solution, DiagNet, is able to build generic models that (i) do not make any assumption on the underlying network topology, (ii) do not require to define the full set of possible causes during training, and (iii) can be quickly adapted to diagnose new services.

DiagNet adapts recent image analysis tactics for system and network metrics, collected from a large and dynamic set of landmark servers. In details, it applies non-overlapping convolutions and global pooling to extract generic information about the analyzed network. This genericness allows to build a general model, that can later be generalized to any Internet service with minimal effort. DiagNet leverages backpropagation attention mechanisms to extend the possible root causes to the set of available metrics, making the model fully extensible. We evaluated DiagNet on geodistributed mockup web services and automated users running in 6 AWS regions, and demonstrated promising root cause analysis capabilities. While this initial work is being reviewed, we are deploying DiagNet for real web services and users to evaluate its performance in a more realistic setup.

Christoph Neumann (InterDigital) actively participated in this work.

6.5.3. Application-aware adaptive partitioning for graph processing systems

Participant: Erwan Le Merrer.

Modern online applications value real-time queries over fresh data models. This is the case for graph-based applications, such as social networking or recommender systems,running on front-end servers in production. A core problem in graph processing systems is the efficient partitioning of the input graph over multiple workers. Recent advances over Bulk Synchronous Parallel processing systems (BSP) enabled computations over partitions on those workers, independently of global synchronization supersteps. A good objective partitioning makes the understanding of the load balancing and communication trade-off mandatory for performance improvement. This work [32] addresses this trade-off through the proposal of an optimization problem, that is to be solved continuously to avoid performance degradation over time. Our simulations show that the design of the software module we propose yields significant performance improvements over the BSP processing model.

This work was done in collaboration with Gilles Trédan (LAAS/CRNS).

6.5.4. How to Spread a Rumor: Call Your Neighbors or Take a Walk?

Participant: George Giakkoupis.

In [28], we study the problem of randomized information dissemination in networks. We compare the now standard push-pull protocol, with agent-based alternatives where information is disseminated by a collection of agents performing independent random walks. In the visit-exchange protocol, both nodes and agents store information, and each time an agent visits a node, the two exchange all the information they have. In the meet-exchange protocol, only the agents store information, and exchange their information with each agent they meet.

We consider the broadcast time of a single piece of information in an n-node graph for the above three protocols, assuming a linear number of agents that start from the stationary distribution. We observe that there are graphs on which the agent-based protocols are significantly faster than push-pull, and graphs where the converse is true. We attribute the good performance of agent-based algorithms to their inherently fair bandwidth utilization, and conclude that, in certain settings, agent-based information dissemination, separately or in combination with push-pull, can significantly improve the broadcast time.

The graphs considered above are highly non-regular. Our main technical result is that on any regular graph of at least logarithmic degree, push-pull and visit-exchange have the same asymptotic broadcast time. The proof uses a novel coupling argument which relates the random choices of vertices in push-pull with the random walks in visit-exchange. Further, we show that the broadcast time of meet-exchange is asymptotically at least as large as the other two's on all regular graphs, and strictly larger on some regular graphs.

As far as we know, this is the first systematic and thorough comparison of the running times of these very natural information dissemination protocols.

This work was done in collaboration with Frederik Mallmann-Trenn (MIT) and Hayk Saribekyan (University of Cambridge, UK).

HYBRID Project-Team

7. New Results

7.1. Virtual Reality Tools and Usages

7.1.1. Studying the Mental Effort in Virtual Versus Real Environments

Participants: Tiffany Luong, Ferran Argelaguet, Anatole Lécuyer [contact].

Is there an effect of Virtual Reality (VR) Head-Mounted Display (HMD) on the user's mental effort? In this work, we compare the mental effort in VR versus in real environments [26]. An experiment (N=27) was conducted to assess the effect of being immersed in a virtual environment (VE) using a HMD on the user's mental effort while performing a standardized cognitive task (the wellknown N-back task, with three levels of difficulty (1,2,3)). In addition to test the effect of the environment (i.e., virtual versus real), we also explored the impact of performing a dual task (i.e., sitting versus walking) in both environments on mental effort. The mental effort was assessed through self-reports, task performance, behavioural and physiological measures. In a nutshell, the analysis of all measurements revealed no significant effect of being immersed in the VE on the users' mental effort. In contrast, natural walking significantly increased the users' mental effort. Taken together, our results support the fact that there is no specific additional mental effort related to the immersion in a VE using a VR HMD.

7.1.2. Influence of Personality Traits and Body Awareness on the Sense of Embodiment in VR

Participants: Diane Dewez, Rebecca Fribourg, Ferran Argelaguet, Anatole Lécuyer [contact].

With the increasing use of avatars in virtual reality, it is important to identify the factors eliciting the sense of embodiment. This work reports an exploratory study aiming at identifying internal factors (personality traits and body awareness) that might cause either a resistance or a predisposition to feel a sense of embodiment towards a virtual avatar. To this purpose, we conducted an experiment (n=123) in which participants were immersed in a virtual environment and embodied in a gender-matched generic virtual avatar through a head-mounted display [16]. After an exposure phase in which they had to perform a number of visuomotor tasks, a virtual character entered the virtual scene and stabbed the participants' virtual hand with a knife (see Figure 4). The participants' sense of embodiment was measured, as well as several personality traits (Big Five traits and locus of control) and body awareness, to evaluate the influence of participants' personality on the acceptance of the virtual body. The major finding is that the locus of control is linked to several components of embodiment: the sense of agency is positively correlated with an internal locus of control and the sense of body ownership is positively correlated with an external locus of control. Taken together, our results suggest that the locus of control could be a good predictor of the sense of embodiment. Yet, further studies are required to confirm these results.
This work was done in collaboration with the MimeTIC team.

7.1.3. Consumer perceptions and purchase behavior of imperfect fruits and vegetables in VR

Participants: Jean-Marie Normand, Guillaume Moreau [contact].

This study investigates the effects of fruits and vegetables (FaVs) abnormality on consumer perceptions and purchasing behavior [9]. For the purposes of this study, a virtual grocery store was created with a fresh FaVs section, where 142 participants became immersed using an Oculus Rift DK2 Head-Mounted Display (HMD) software. Participants were presented either normal, slightly misshapen, moderately misshapen or severely misshapen FaVs. The study findings indicate that shoppers tend to purchase a similar number of FaVs whatever their level of deformity. However, perceptions of the appearance and quality of the FaVs depend on the degree of abnormality. Moderately misshapen FaVs are perceived as significantly better than those that are heavily misshapen but also "slightly" misshapen (except for the appearance of fruits).

Figure 4. From left to right: an example of a trajectory to draw during the experimental task; A view of the scene from behind; Another virtual character stabbing the participants' virtual hand at the end of the experiment to measure their response to the threat on their virtual body.

This work was done in collaboration with Audecia Recherce, the University of Reading and the University of Tokyo.

7.1.4. Am I better in VR with a real audience?

Participants: Romain Terrier, Valérie Gouranton [contact], Bruno Arnaldi.

We designed an experimental study to investigate the effects of a real audience on social inhibition [33]. The study is a virtual reality (VR) and multiuser application (see Figure 5). The experience is locally or remotely shared. The application engages one user and a real audience (i.e., local or remote conditions). A control condition is designed where the user is alone (i.e., alone condition). The objective performance (i.e., type and answering time) of users, when performing a categorization of numbers task in VR, is used to explore differences between conditions. In addition to this, the perceptions of others, the stress, the cognitive workload, and the presence of each user have been compared in relation to the location of the real audience. The results showed that in the presence of a real audience (in the local and remote conditions), user performance is affected by social inhibitions. Furthermore, users are even more influenced when the audience does not share the same room, despite others are less perceived.
This work was done in collaboration with IRT B COM.

Figure 5. Experimental setup for the social inhibition experiment in Virtual Reality.

7.1.5. Create by Doing – Action sequencing in VR

Participants: Flavien Lécuyer, Valérie Gouranton [contact], Adrien Reuzeau, Ronan Gaugne, Bruno Arnaldi.

In every virtual reality application, there are actions to perform, often in a logical order. This logical ordering can be a predefined sequence of actions, enriched with the representation of different possibilities, which we refer to as a scenario. Authoring such a scenario for virtual reality is still a difficult task, as it needs both the expertise from the domain expert and the developer. We propose [28] to let the domain expert create in virtual reality the scenario by herself without coding, through the paradigm of creating by doing (see Figure 6). The domain expert can run an application, record the sequence of actions as a scenario, and then reuse this scenario for other purposes, such as an automatic replay of the scenario by a virtual actor to check the obtained scenario, the injection of this scenario as a constraint or a guide for a trainee, or the monitoring of the scenario unfolding during a procedure.

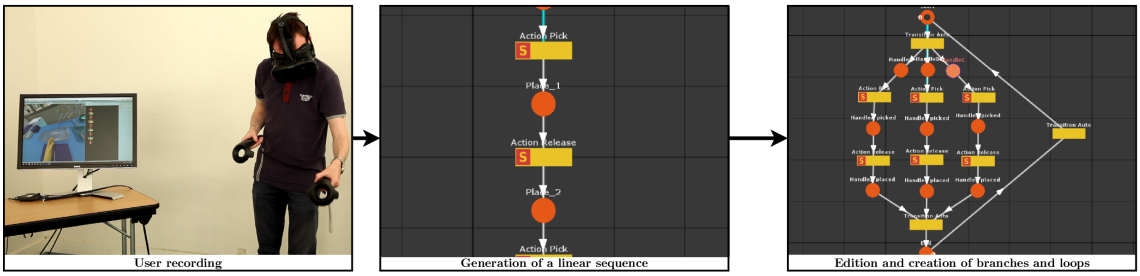


Figure 6. The proposed workflow for the creation of scenarios

7.1.6. Help! I Need a Remote Guide in my Mixed Reality Collaborative Environment

Participants: Valérie Gouranton [contact], Bruno Arnaldi.

The help of a remote expert in performing a maintenance task can be useful in many situations, and can save time as well as money. In this context, augmented reality (AR) technologies can improve remote guidance thanks to the direct overlay of 3D information onto the real world. Furthermore, virtual reality (VR) enables a remote expert to virtually share the place in which the physical maintenance is being carried out. In a traditional local collaboration, collaborators are face-to-face and are observing the same artifact, while being able to communicate verbally and use body language such as gaze direction or facial expression. These interpersonal communication cues are usually limited in remote collaborative maintenance scenarios, in which the agent uses an AR setup while the remote expert uses VR. Providing users with adapted interaction and awareness features to compensate for the lack of essential communication signals is therefore a real challenge for remote MR collaboration. However, this context offers new opportunities for augmenting collaborative abilities, such as sharing an identical point of view, which is not possible in real life. Based on the current task of the maintenance procedure, such as navigation to the correct location or physical manipulation, the remote expert may choose to freely control his/her own viewpoint of the distant workspace, or instead may need to share the viewpoint of the agent in order to better understand the current situation. In this work, we first focus on the navigation task, which is essential to complete the diagnostic phase and to begin the maintenance task in the correct location [8]. We then present a novel interaction paradigm, implemented in an early prototype, in which the guide can show the operator the manipulation gestures required to achieve a physical task that is necessary to perform the maintenance procedure. These concepts are evaluated, allowing us to provide guidelines for future systems targeting efficient remote collaboration in MR environments.
This work was done in collaboration with IRT B COM and UMR Lab-STICC, France.

7.1.7. Learning procedural skills with a VR simulator: An acceptability study

Participants: Valérie Gouranton [contact], Bruno Arnaldi.

Virtual Reality (VR) simulation has recently been developed and has improved surgical training. Most VR simulators focus on learning technical skills and few on procedural skills. Studies that evaluated VR simulators focused on feasibility, reliability or easiness of use, but few of them used a specific acceptability measurement tool. The aim of the study was to assess acceptability and usability of a new VR simulator for procedural skill training among scrub nurses, based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model. The simulator training system was tested with a convenience sample of 16 non-expert users and 13 expert scrub nurses from the neurosurgery department of a French University Hospital. The scenario was designed to train scrub nurses in the preparation of the instrumentation table for a craniotomy in the operating room (OR). Acceptability of the VR simulator was demonstrated with no significant difference between expert scrub nurses and non-experts. There was no effect of age, gender or expertise. Workload, immersion and simulator sickness were also rated equally by all participants. Most participants stressed its pedagogical interest, fun and realism, but some of them also regretted its lack of visual comfort. This VR simulator designed to teach surgical procedures can be widely used as a tool in initial or vocational training [2], [43].
This work was achieved in collaboration with Univ. Rennes 2-LP3C, LTSI and the Hycomes team.

7.1.8. The Anisotropy of Distance Perception in VR

Participants: Etienne Peillard, Anatole Lécuyer, Ferran Argelaguet, Jean-Marie Normand, Guillaume Moreau [contact].

The topic of distance perception has been widely investigated in Virtual Reality (VR). However, the vast majority of previous work mainly focused on distance perception of objects placed in front of the observer. Then, what happens when the observer looks on the side? In this work, we study differences in distance estimation when comparing objects placed in front of the observer with objects placed on his side [31]. Through a series of four experiments (n=85), we assessed participants' distance estimation and ruled out potential biases. In particular, we considered the placement of visual stimuli in the field of view, users' exploration behavior as well as the presence of depth cues. For all experiments a two-alternative forced choice (2AFC) standardized psychophysical protocol was employed, in which the main task was to determine the stimuli that seemed to be the farthest one. In summary, our results showed that the orientation of virtual stimuli with respect to the user introduces a distance perception bias: objects placed on the sides are systematically perceived farther away than objects in front. In addition, we could observe that this bias increases along with the angle, and appears to be independent of both the position of the object in the field of view as well as the quality of the virtual scene. This work sheds a new light on one of the specificities of VR environments regarding the wider subject of visual space theory. Our study paves the way for future experiments evaluating the anisotropy of distance perception in real and virtual environments.

7.1.9. Study of Gaze and Body Segments Temporal Reorientation Behaviour in VR

Participants: Hugo Brument, Ferran Argelaguet [contact].

This work investigates whether the body anticipation synergies in real environments (REs) are preserved during navigation in virtual environments (VEs). Experimental studies related to the control of human locomotion in REs during curved trajectories report a top-down body segments reorientation strategy, with the reorientation of the gaze anticipating the reorientation of head, the shoulders and finally the global body motion [12]. This anticipation behavior provides a stable reference frame to the walker to control and reorient his/her body segments according to the future walking direction. To assess body anticipation during navigation in VEs, we conducted an experiment where participants, wearing a head-mounted display, performed a lemniscate trajectory in a virtual environment (VE) using five different navigation techniques, including walking, virtual steering (head, hand or torso steering) and passive navigation. For the purpose of this experiment, we designed a new control law based on the power-law relation between speed and curvature during human walking. Taken together, our results showed a similar ordered top-down sequence of reorientation of the gaze, head and shoulders during curved trajectories for all the evaluated techniques. However, the anticipation mechanism was significantly higher for the walking condition compared to the

others. Finally, the results work pave the way to the better understanding of the underlying mechanisms of human navigation in VEs and to the design of navigation techniques more adapted to humans.

This work was done in collaboration with the MimeTIC team and the Interactive Media Systems Group (TU Wien, Vienna, Austria).

7.1.10. User-centered design of a multisensory power wheelchair simulator

Participants: Guillaume Vailland, Valérie Gouranton [contact].

Autonomy and social inclusion can reveal themselves everyday challenges for people experiencing mobility impairments. These people can benefit from technical aids such as power wheelchairs to access mobility and overcome social exclusion. However, power wheelchair driving is a challenging task which requires good visual, cognitive and visuo-spatial abilities. Besides, a power wheelchair can cause material damage or represent a danger of injury for others or oneself if not operated safely. Therefore, training and repeated practice are mandatory to acquire safe driving skills to obtain power wheelchair prescription from therapists. However, conventional training programs may reveal themselves insufficient for some people with severe impairments. In this context, Virtual Reality offers the opportunity to design innovative learning and training programs while providing realistic wheelchair driving experience within a virtual environment. In line with this, we propose a user-centered design of a multisensory power wheelchair simulator [34]. This simulator addresses classical virtual experience drawbacks such as cybersickness and sense of presence by combining 3D visual rendering, haptic feedback and motion cues. The simulator was showcased in the SOFMER conference [37].
This work has been done in collaboration with Rainbow team.

Figure 7. Wheelchair simulator.

7.1.11. Machine Learning Based Interaction Technique Selection For 3D User Interfaces

Participant: Bruno Arnaldi [contact].

A 3D user interface can be adapted in multiple ways according to each user's needs, skills and preferences. Such adaptation can consist in changing the user interface layout or its interaction techniques. Personalization systems which are based on user models can automatically determine the configuration of a 3D user interface in order to fit a particular user. In this work, we proposed to explore the use of machine learning in order to propose a 3D selection interaction technique adapted to a target user [23]. To do so, we built a dataset with 51 users on a simple selection application in which we recorded each user profile, his/her results to a

2D Fitts Law based pre-test and his/her preferences and performances on this application for three different interaction techniques. Our machine learning algorithm based on Support Vector Machines (SVMs) trained on this dataset proposes the most adapted interaction technique according to the user profile or his/her result to the 2D selection pre-test. Our results suggest the interest of our approach for personalizing a 3D user interface according to the target user but it would require a larger dataset in order to increase the confidence about the proposed adaptations.

7.1.12. The 3DUI Contest 2019

Participants: Hugo Brument, Rebecca Fribourg, Gerard Gallagher, Thomas Howard, Flavien Lécuyer, Tiffany Luong, Victor Mercado, Etienne Peillard, Xavier de Tinguy, Maud Marchal [contact].

Pyramid Escape: Design of Novel Passive Haptics Interactions for an Immersive and Modular Scenario
In this work, we present the design of ten different 3D user interactions using passive haptics and embedded in an escape game scenario in which users have to escape from a pyramid in a limited time [11]. Our solution is innovative by its modularity, allowing interactions with virtual objects using tangible props manipulated either directly using the hands and feet or indirectly through a single prop held in the hand, in order to perform several interactions with the virtual environment (VE). We also propose a navigation technique based on the "impossible spaces" design, allowing users to naturally walk through several overlapping rooms of the VE. All together, our different interaction techniques allow the users to solve several enigmas built into a challenging scenario inside a pyramid.

7.2. Augmented Reality Tools and Usages

7.2.1. Authoring AR by AR, abstraction and libraries

Participants: Flavien Lécuyer, Valérie Gouranton [contact], Adrien Reuzeau, Ronan Gaugne, Bruno Arnaldi.

The demand for augmented reality applications is rapidly growing. In many domains, we observe a new interest for this technology, stressing the need for more efficient ways of producing augmented content. Similarly to virtual reality, interactive objects in augmented reality are a powerful means to improve the experience. While it is now well democratized for virtual reality, interactivity is still finding its way into augmented reality. To open the way to this interactive augmented reality, we designed a new methodology for the management of the interactions in augmented reality, supported by an authoring tool for the use by designers and domain experts [27]. This tool makes the production of interactive augmented content faster, while being scalable to the needs of each application. Usually in the creation of applications, a large amount of time is spent through discussions between the designer (or the domain expert), carrying the needs of the application, and the developer, holding the knowledge to create it (see Figure 8). Thanks to our tool, we reduce this time by allowing the designer to create an interactive application, without having to write a single line of code.

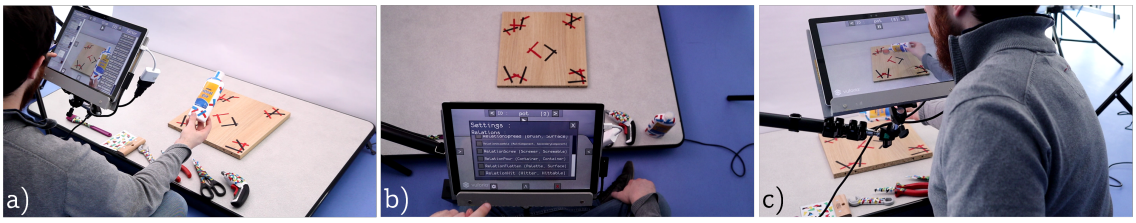


Figure 8. From left to right, the user (a) adds an interactive behaviour on a bottle of glue, (b) imports the interactions in the environment, and (c) uses the interaction to pour the virtual glue from the real bottle into a virtual pot

7.2.2. Studying Exocentric Distance Perception in Optical See-Through AR

Participants: Etienne Peillard, Ferran Argelaguet, Jean-Marie Normand, Anatole Lécuyer, Guillaume Moreau [contact].

While perceptual biases have been widely investigated in Virtual Reality (VR), very few studies have considered the challenging environment of Optical See-through Augmented Reality (OST-AR). Moreover, regarding distance perception, existing works mainly focus on the assessment of egocentric distance perception, i.e. distance between the observer and a real or a virtual object. In this work, we studied exocentric distance perception in AR, hereby considered as the distance between two objects, none of them being directly linked to the user. We report a user study (n=29) aiming at estimating distances between two objects lying in a frontoparallel plane at 2.1m from the observer (i.e. in the medium-field perceptual space). Four conditions were tested in our study: real objects on the left and on the right of the participant (called real-real), virtual objects on both sides (virtual-virtual), a real object on the left and a virtual one on the right (real-virtual) and finally a virtual object on the left and a real object on the right (virtual-real). Participants had to reproduce the distance between the objects by spreading two real identical objects presented in front of them (see Figure 9). The main findings of this study are the overestimation (20%) of exocentric distances for all tested conditions. Surprisingly, the real-real condition was significantly more overestimated (by about 4%,p=.0166) compared to the virtual-virtual condition, i.e. participants obtained better estimates of the exocentric distance for the virtual-virtual condition. Finally, for the virtual-real/real-virtual conditions, the analysis showed a non-symmetrical behavior, which suggests that the relationship between real and virtual objects with respect to the user might be affected by other external factors. Considered together, these unexpected results illustrate the need for additional experiments to better understand the perceptual phenomena involved in exocentric distance perception with real and virtual objects [30].

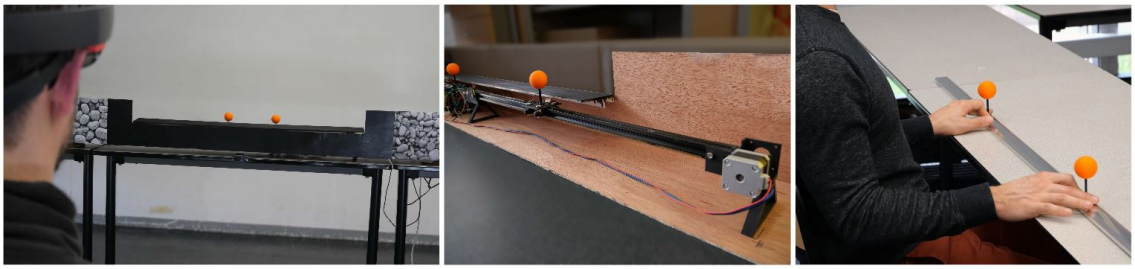


Figure 9. Left, bench displaying two real spheres. The hinge-actuated moving panel, opened here, could be automatically opened/closed to reveal/hide the visual stimuli. Center, one of the two rails of the bench, seen from behind. An orange sphere is attached on top of a trolley that can slide on the rail. The trolley is moved by a stepper motor through a belt. The other half of the bench is symmetrical. Right, participants could provide the perceived exocentric distance by placing two sliding spheres. After the participants placed the spheres the system automatically took a picture of both spheres which was used to measure the distance between both spheres.

7.2.3. Influence of virtual objects' shadows and lighting coherence in AR

Participants: Etienne Peillard, Jean-Marie Normand, Guillaume Moreau [contact].

This work focuses on how virtual objects' shadows as well as differences in alignment between virtual and real lighting influence distance perception in optical see-through (OST) augmented reality (AR) [5]. Four hypotheses are pro-posed: (H1) Participants underestimate distances in OST AR; (H2) Virtual objects' shadows improve distance judgment accuracy in OST AR; (H3) Shadows with different realism levels have different influence on distance perception in OST AR; (H4) Different levels of lighting misalignment between

real and virtual lights have different influence on distance perception in OST AR scenes. Two experiments were designed with an OST head mounted display(HMD), the Microsoft HoloLens. Participants had to match the position of a virtual object displayed in the OST-HMD with a real target. Distance judgment accuracy was recorded under the different shadows and lighting conditions.The results validate hypotheses H2 and H4 but surprisingly showed no impact of the shape of virtual shadows on distance judgment accuracy thus rejecting hypothesis H3. Regarding hypothesis H1, we detected a trend toward underestimation; given the high variance of the data, more experiments are needed to confirm this result. Moreover, the study also reveals that perceived distance errors and completion time of trials increase along with targets' distance.

7.2.4. A study on differences in human perception in AR

Participants: Jean-Marie Normand, Guillaume Moreau [contact].

With the recent growth in the development of augmented reality (AR) technologies, it is becoming important to study human perception of AR scenes. In order to detect whether users will suffer more from visual and operator fatigue when watching virtual objects through optical see-through head-mounted displays (OST-HMDs), compared with watching real objects in the real world, we propose a comparative experiment including a virtual magic cube task and a real magic cube task [4]. The scores of the subjective questionnaires (SQ) and the values of the critical flicker frequency (CFF) were obtained from 18 participants. In our study, we use several electrooculogram (EOG) and heart rate variability (HRV) measures as objective indicators of visual and operator fatigue. Statistical analyses were performed to deal with the subjective and objective indicators in the two tasks. Our results suggest that participants were very likely to suffer more from visual and operator fatigue when watching virtual objects presented by the OST-HMD. In addition, the present study provides hints that HRV and EOG measures could be used to explore how visual and operator fatigue are induced by AR content. Finally, three novel HRV measures are proposed to be used as potential indicators of operator fatigue.
This work was done in collaboration with the Beijing Engineering Research Center of Mixed Reality and Advanced Display (School of Optics and Photonics, Beijing Institute of Technology, Beijing, China) and AICFVE (Beijing Film Academy, Beijing, China).

7.3. Physically-Based Simulation and Haptic Feedback

7.3.1. Design of haptic guides for pre-positioning assistance of a comanipulated needle

Participant: Maud Marchal [contact].

In minimally-invasive procedures like biopsy, the physician has to insert a needle into the tissues of a patient to reach a target. Currently, this task is mostly performed manually and under visual guidance. However, manual needle insertion can result in a large final positioning error of the tip that might lead to misdiagnosis and inadequate treatment. A way to solve this limitation is to use shared control; a gesture assistance paradigm that combines the cognitive skills of the operator with the precision, stamina and repeatability of a robotic or haptic device. In this paper, we propose to assist the physician with a haptic device that holds the needle and generates mechanical guides during the phase of manual needle pre-positioning. In the latter, the physician has to place the tip of the needle on a planned entry point, with a pre-defined angle of incidence. From this pre-operative information and also from intra-operative measurements, we propose to generate haptic cues, known as virtual fixtures, to guide the physician towards the desired position and orientation of the needle. It takes the form of five haptic guides, each one implementing virtual fixtures. We conducted a user study where those guides were compared to the unassisted reference gesture. The most constraining guide, in terms of assisted degrees of freedom, was highlighted as the one that provides the best results in terms of performance and user experience [20], [21].
This work was done in collaboration with the Inria Rainbow team.

7.3.2. An Interactive Physically-based Model for Active Suction Phenomenon Simulation

Participants: Antonin Bernardin, Maud Marchal [contact].

While suction cups are widely used in Robotics, the literature is underdeveloped when it comes to the modelling and simulation of the suction phenomenon (see Figure 10). In this work, we present a novel physically-based approach to simulate the behavior of active suction cups. Our model relies on a novel formulation which assumes the pressure exerted on a suction cup during active control is based on constraint resolution. Our algorithmic implementation uses a classification process to handle the contacts during the suction phenomenon of the suction cup on a surface. Then, we formulate a convenient way for coupling the pressure constraint with the multiple contact constraints. We propose an evaluation of our approach through a comparison with real data, showing the ability of our model to reproduce the behavior of suction cups. Our approach paves the way for improving the design as well as the control of robotic actuators based on suction cups such as vacuum grippers.

This work was done in collaboration with the Inria Defrost team.

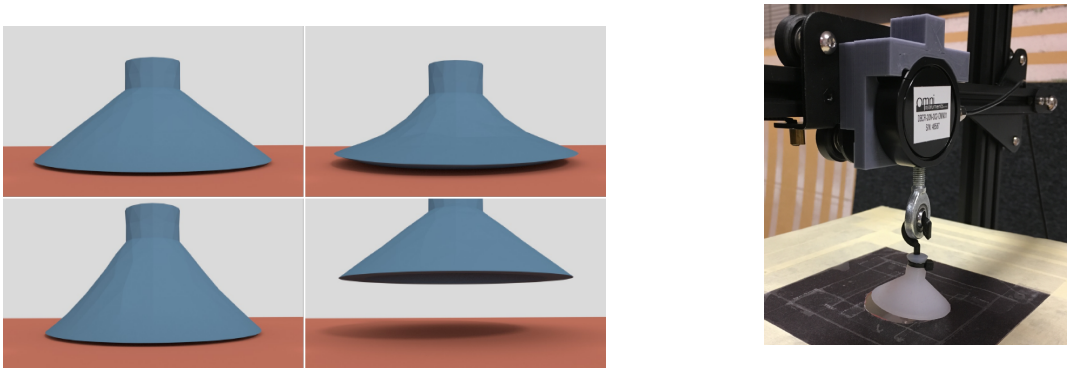


Figure 10. Left, Illustration of our constraint-based physically-based approach for simulating active suction cup phenomenon: (Top) the suction cup is actively stuck to the surface, (Bottom) is is then release until being completely in the air. Right, experimental setup for the force measurements. The suction cup is attached to a force sensor. When it is positioned on a flat surface, its cavity is linked to a vacuum pump with a regulator inbetween.

7.3.3. How different tangible and virtual objects can be while still feeling the same?

Participants: Xavier de Tinguy, Anatole Lécuyer, Maud Marchal [contact].

Tangible objects are used in Virtual Reality to provide human users with distributed haptic sensations when grasping virtual objects. To achieve a compelling illusion, there should be a good correspondence between the haptic features of the tangible object and those of the corresponding virtual one, i.e., what users see in the virtual environment should match as much as possible what they touch in the real world. This work [14] aims at quantifying how similar tangible and virtual objects need to be, in terms of haptic perception, to still feel the same. As it is often not possible to create tangible replicas of all the virtual objects in the scene, it is important to understand how different tangible and virtual objects can be without the user noticing (see Figure 11). This paper reports on the just-noticeable difference (JND) when grasping, with a thumb-index pinch, a tangible object which differ from a seen virtual one on three important haptic features: width, local orientation, and curvature. Results show JND values of 5.75%, 43.8%, and 66.66% of the reference shape for the width, local orientation, and local curvature features, respectively. These results will enable researchers in the field of Virtual Reality to use a reduced number of tangible objects to render multiple virtual ones.
This work was done in collaboration with the Inria Rainbow team.

7.3.4. Toward Universal Tangible Objects

Participants: Xavier de Tinguy, Maud Marchal, Anatole Lécuyer [contact].

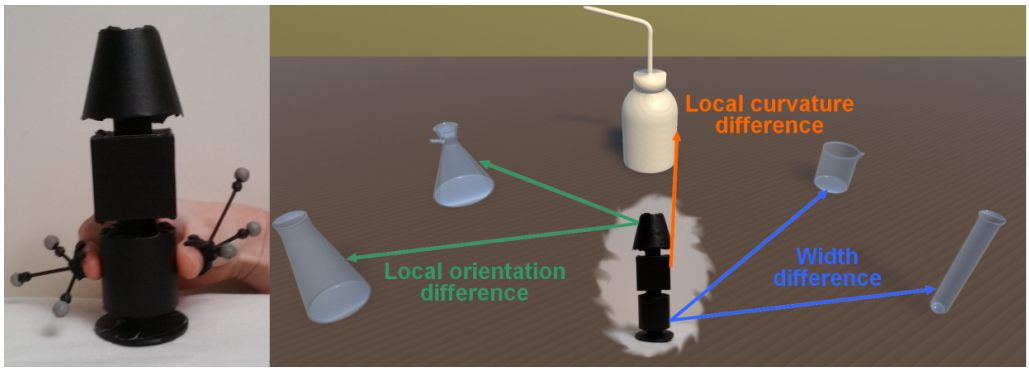


Figure 11. Understanding how different a tangible object (left) can be from virtual objects (right) without the user noticing the mismatch. We focused our study on three specific criteria: width, local orientation, and curvature.

Tangible objects are a simple yet effective way for providing haptic sensations in Virtual Reality. For achieving a compelling illusion, there should be a good correspondence between what users see in the virtual environment and what they touch in the real world. The haptic features of the tangible object should indeed match those of the corresponding virtual one in terms of, e.g., size, local shape, mass, texture. A straightforward solution is to create perfect tangible replicas of all the virtual objects in the scene. However, this is often neither feasible nor desirable. This work [15] presents an innovative approach enabling the use of few tangible objects to render many virtual ones (see Figure 12). The proposed algorithm analyzes the available tangible and virtual objects to find the best grasps in terms of matching haptic sensations. It starts by identifying several suitable pinching poses on the considered tangible and virtual objects. Then, for each pose, it evaluates a series of haptically-salient characteristics. Next, it identifies the two most similar pinching poses according to these metrics, one on the tangible and one on the virtual object. Finally, it highlights the chosen pinching pose, which provides the best matching sensation between what users see and touch. The effectiveness of our approach is evaluated through a user study. Results show that the algorithm is able to well combine several haptically-salient object features to find convincing pinches between the given tangible and virtual objects.
This work was done in collaboration with the Inria Rainbow team.

7.3.5. Investigating the recognition of local shapes using mid-air ultrasound haptics

Participants: Thomas Howard, Gerard Gallagher, Anatole Lécuyer, Maud Marchal [contact].

Mid-air haptics technologies are able to convey haptic sensations without any direct contact between the user and the haptic interface. One representative example of this technology is ultrasound haptics, which uses ultrasonic phased arrays to deliver haptic sensations. Research on ultrasound haptics is only in its beginnings, and the literature still lacks principled perception studies in this domain. This work [22] presents a series of human subject experiments investigating important perceptual aspects related to the rendering of 2D shapes by an ultrasound haptic interface (the Ultrahaptics STRATOS platform, see Figure 13). We carried out four user studies aiming at evaluating (i) the absolute detection threshold for a static focal point rendered via amplitude modulation, (ii) the absolute detection and identification thresholds for line patterns rendered via spatiotemporal modulation, (iii) the ability to discriminate different line orientations, and (iv) the ability to perceive virtual bumps and holes. These results shed light on the rendering capabilities and limitations of this novel technology for 2D shapes.
This work was done in collaboration with the Inria Rainbow team.

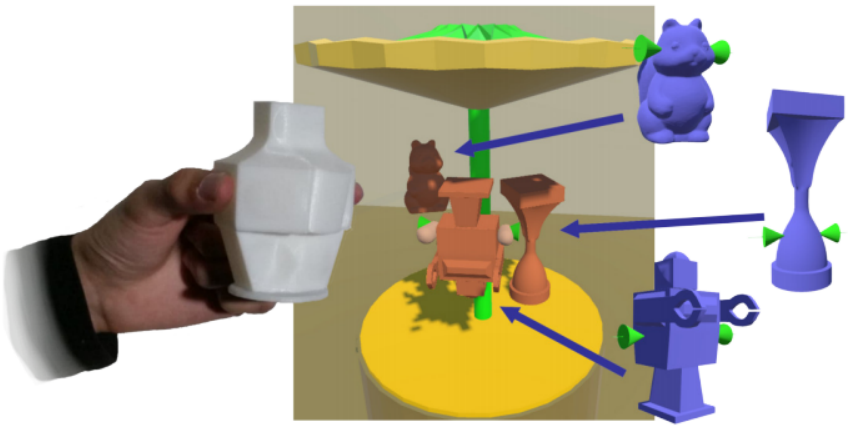


Figure 12. Illustration of our approach through a carousel of virtual objects that can be grasped using a single "universal" tangible object. The user is able to turn the virtual carousel and manipulate the three virtual objects using the suggested pinch poses (in green). These poses are proposed by our algorithm to best match the corresponding haptic pinching sensations on the tangible object.

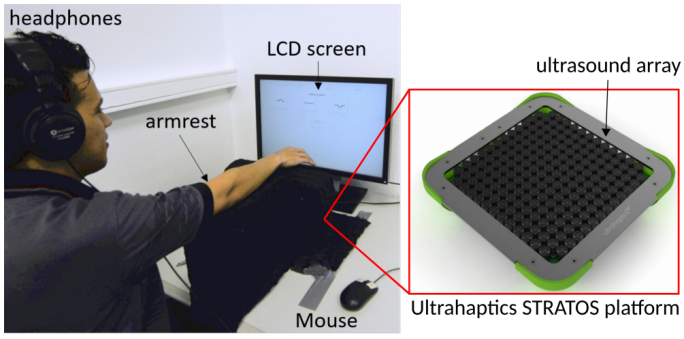


Figure 13. Experimental setup to investigate the recognition of local shapes using mid-air ultrasound haptics.

7.3.6. Touchy: Tactile Sensations on Touchscreens Using a Cursor and Visual Effects

Participants: Antoine Costes, Ferran Argelaguet, Anatole Lécuyer [contact].

Haptic enhancement of touchscreens usually involves vibrating motors that produce limited sensations or custom mechanical actuators that are difficult to widespread. In this work, we propose an alternative approach called "Touchy" to induce haptic sensations in touchscreens through purely visual effects [3]. Touchy introduces a symbolic cursor under the user's finger which shape and motion are altered in order to evoke haptic properties. This novel metaphor enables to address four different perceptual dimensions, namely: hardness, friction, fine roughness and macro roughness. Our metaphor comes with a set of seven visual effects that we compared with real texture samples within a user study conducted with 14 participants. Taken together our results show that Touchy is able to elicit clear and distinct haptic properties: stiffness, roughness, reliefs, stickiness and slipperiness.
This work was achieved in collaboration with InterDigital.

7.3.7. Investigating Tendon Vibration Illusions

Participants: Salomé Lefranc [contact], Mélanie Cogné, Mathis Fleury, Anatole Lécuyer.

Illusion of movement induced by tendon vibration can be useful in applications such as rehabilitation of neurological impairments. In [40], we investigated whether a haptic proprioceptive illusion induced by a tendon vibration of the wrist congruent to the visual feedback of a moving hand could increase the overall illusion of movement. Tendon vibration was applied on the non-dominant wrist during 3 visual conditions: a moving virtual hand corresponding to the movement that the subjects could feel during the tendon vibration (Moving condition), a static virtual hand (Static condition), or no virtual hand at all (Hidden condition). There was a significant difference between the 3 visual feedback conditions, and the Moving condition was found to induce a higher intensity of illusion of movement and higher sensation of wrist's extension. Therefore, our study demonstrated the potentiation of illusion by visual cues congruent to the illusion of movement. Further steps will be to test the same hypothesis with stroke patients and use our results to develop EEG-based Neurofeedback including vibratory feedback to improve upper limb motor function after a stroke.
This work was achieved in collaboration with CHU Rennes and Inria EMPENN team.

7.4. Brain-Computer Interfaces

7.4.1. Defining Brain-Computer Interfaces: A Human-Computer Interaction Perspective

Participants: Hakim Si Mohammed, Ferran Argelaguet, Anatole Lécuyer [contact].

Regardless of the term used to designate them, Brain-Computer Interfaces (BCIs) are "Interfaces" between a user and a computer in the broad sense of the term. This paper aims to discuss how BCIs have been defined in the literature from the day the term was introduced by Jacques Vidal. In [32], from a Human-Computer Interaction perspective, we propose a new definition of Brain-Computer Interfaces as : "any artificial systems that directly converts brain activity into input of a computer process". As they are interfaces, such definition should not include the finality and objective of the system they are used to interact with. To illustrate this, we compared BCIs with other widely used Human-Computer Interfaces, and drew analogies in their conception and purpose.

This work was done in collaboration with the Inria LOKI team.

7.4.2. A conceptual space for EEG-based brain-computer interfaces

Participant: Anatole Lécuyer [contact].

Brain-Computer Interfaces have become more and more popular these last years. Researchers use this technology for several types of applications, including attention and workload measures but also for the direct control of objects by the means of BCIs. In [7] we present a first, multidimensional feature space for EEG-based BCI applications to help practitioners to characterize, compare and design systems, which use EEG-based BCIs. Our feature space contains 4 axes and 9 sub-axes and consists of 41 options in total as well as their different combinations. In addition we present the axes of our feature space and we position our feature space regarding the existing BCI and HCI taxonomies. We also showed how our work integrates the past works, and/or complements them.

7.4.3. The use of haptic feedback in Brain-Computer Interfaces and Neurofeedback

Participants: Mathis Fleury, Anatole Lécuyer [contact].

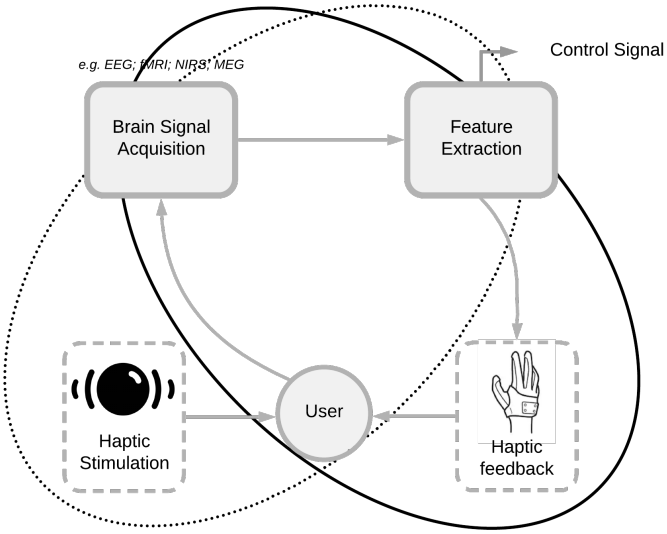


Figure 14. Using haptic feedback in active and reactive Brain-Computer Interfaces (BCI). In active BCI, haptics provide feedback from user's neural activity (black ellipse). In reactive BCI, haptics provide a stimulation to elicit a specific brain activity (black doted ellipse).

Neurofeedback (NF) and brain-computer interfaces are based on the recording of the cerebral activity associated with the requested task and the presentation of a feedback. The subject relies on the given feedback (visual, auditory or haptic) to learn and improve his mental strategy. It is therefore of crucial importance that it must be transmitted optimally. Historically, vision is the most used sensory modality in BCI/NF applications, but its use is raising potential issues. The more and more frequent use of haptic as a feedback modality reveals the limits of visual feedback; indeed, a visual feedback is not suitable in some cases, for individuals with an impaired visual system or during a mental motor imagery task (e.g. requiring a great abstraction). In such case, a haptic feedback would seem more appropriate. Haptic feedback has also been reported to be more engaging than visual feedback. This feedback could also contribute to close the sensory-motor loop. Haptic-based BCI/NF is a promising alternative for the design of the feedback and potentially improve the clinical efficacy of NF. In [38], [39] we have therefore surveyed the recent studies exploiting haptic feedback in BCI and NF.

This work was achieved in collaboration with the Inria EMPENN team.

7.4.4. Efficacy of EEG-fMRI Neurofeedback for stroke rehabilitation: a pilot study

Participants: Giulia Lioi, Mathis Fleury, Anatole Lécuyer [contact].

Recent studies have shown the potential of neurofeedback for motor rehabilitation after stroke. The majority of these NF approaches have relied solely on one imaging technique: mostly on EEG recordings. Recent study have gone further, revealing the potential of integrating complementary techniques such as EEG and fMRI to achieve a more specific regulation. In this exploratory work, multi-session bimodal EEG-fMRI NF for upper limb motor recovery was tested in four stroke patients.The feasibility of the NF training was investigated [41] with respect to the integrity of the cortico-spinal tract (CST), a well-established predictor of the potential for

clinical improvement. Results indicated that patients exhibiting a high degree of integrity of the ipsilesional CST showed significant increased activation of the ipsilesional M1 at the end of the training. These preliminary findings confirm the critical role of the CST integrity for stroke motor recovery and indicate that this is importantly related also to functional brain regulation of the ipsilesional motor cortex.
This work was achieved in collaboration with Inria EMPENN team.

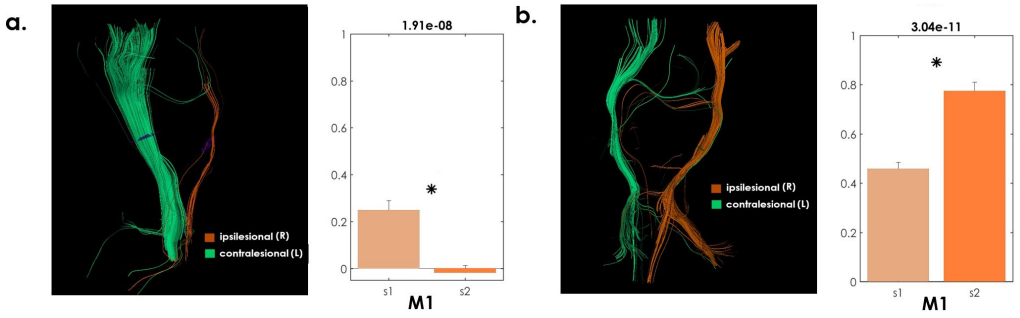


Figure 15. Example of CST reconstruction and primary motor cortex (M1) activation in two patients (a. and b.). Ipsilesional CST is plotted in orange and controlesional CST in green. The bar plot on the right hand side of the figure show the average (and standard error across NF training blocks) of BOLD contrast activation in the primary motor cortex in the first (s1) and second (s2) training session, with relative statistics.

7.4.5. A multi-target motor imagery training using EEG-fMRI Neurofeedback

Participants: Giulia Lioi, Mathis Fleury, Anatole Lécuyer [contact].

Upper limb recovery after stroke is a complex process. Recent studies have revealed the potential of neurofeedback training as an alternative or an aid to traditional therapies. Studies on cerebral plasticity and recovery after stroke indicate that premotor areas should be a preferred target for NF in the most severe patients while M1 stimulation may be more ef for patients with better recovery potential. Moreover, fMRI-NF studies (also on stroke patients) have shown that SMA is a robust correlate of motor imagery, while the activation of M1 is more dif to achieve, especially for short training sessions. Based on these results, in an exploratory work [13], we tested a dynamic NF training more strongly rewarding SMA activation in the NF training session and then increasing the M1 activation contribution in the NF session. We tested this novel approach on four stroke patients in a multisession bimodal EEG-fMRI NF training. To this end, we used an adaptive cortical region of interest (ROI) equal to a weighted combination of ipsilesional SMA and M1 activities and then varied the weights in order guide the patient training towards an improved activation of M1. Four chronic stroke patients with left hemiparesis participated to the study. The experimental protocol included an alternation of bimodal EEG-fMRI NF and unimodal EEG-only NF sessions. Preliminary results, on a short training duration, reveal the potential of a dynamic, multi-target/multimodal NF training approach.
This work was achieved in collaboration with Inria EMPENN team.

7.4.6. Bimodal EEG-fMRI Neurofeedback for upper motor limb rehabilitation

Participants: Giulia Lioi, Mathis Fleury, Anatole Lécuyer [contact].

There is a growing interest in Neurofeedback or Brain computer interfaces for stroke rehabilitation. Integrating EEG and fMRI, two highly complementary imaging modalities, has potential to provide a more specific and efficient stimulation of motor areas. In this exploratory work [25], we tested the feasibility of a multi-session EEG-fMRI NF protocol on four chronic stroke patients, and its potential for upper-limb recovery. All the patients were able to upregulate their activity during NF training with respect to rest in the ipsilesional SMA and M1. Three over four patients showed a significant increase in ipsilesional M1 activation at the end of the protocol. Of these three individuals, two exhibited an increase in FMA-UE score. Preliminary results from this pilot study showed feasibility of bimodal EEG-fMRI in chronic stroke patients and indicated the potential of this training protocol for upper-limb recovery.
This work was achieved in collaboration with Inria EMPENN team.

7.5. Cultural Heritage

7.5.1. Expressive potentials of motion capture in the Vis Insita musical performance

Participants: Ronan Gaugne [contact], Florian Nouviale, Valérie Gouranton.

The electronic music performance project Vis Insita [10] implements the design of experimental instrumental interfaces based on optical motion capture technology with passive infrared markers (MoCap), and the analysis of their use in a real scenic presentation context (Figure 16). Because of MoCap's predisposition to capture the movements of the body, a lot of research and musical applications in the performing arts concern dance or the sonification of gesture. For our research, we wanted to move away from the capture of the human body to analyse the possibilities of a kinetic object handled by a performer, both in terms of musical expression, but also in the broader context of a multimodal scenic interpretation.
This work was done in collaboration with Univ. Rennes 2, France.

Figure 16. The Vis Insita performance.

7.5.2. Interactive and Immersive Tools for Point Clouds in Archaeology

Participants: Ronan Gaugne [contact], Quentin Petit, Valérie Gouranton.

A framework is presented for an immersive and interactive 3D manipulation of large point clouds, in the context of an archaeological study [19]. The framework was designed in an interdisciplinary collaboration with archaeologists. We first applied this framework for the study of an 17th-century building of a Real Tennis court (Figure 17). We propose a display infrastructure associated with a set of tools that allows archaeologists to interact directly with the point cloud within their study process. The resulting framework allows an immersive navigation at scale 1:1 in a dense point cloud, the manipulation and production of cut plans and cross sections, and the positioning and visualisation of photographic views. We also apply the same framework to three other archaeological contexts with different purposes, a 13th century ruined chapel, a 19th-century wreck and a cremation urn from the Iron Age.
This work was done in collaboration with UMR CREAAH, France.

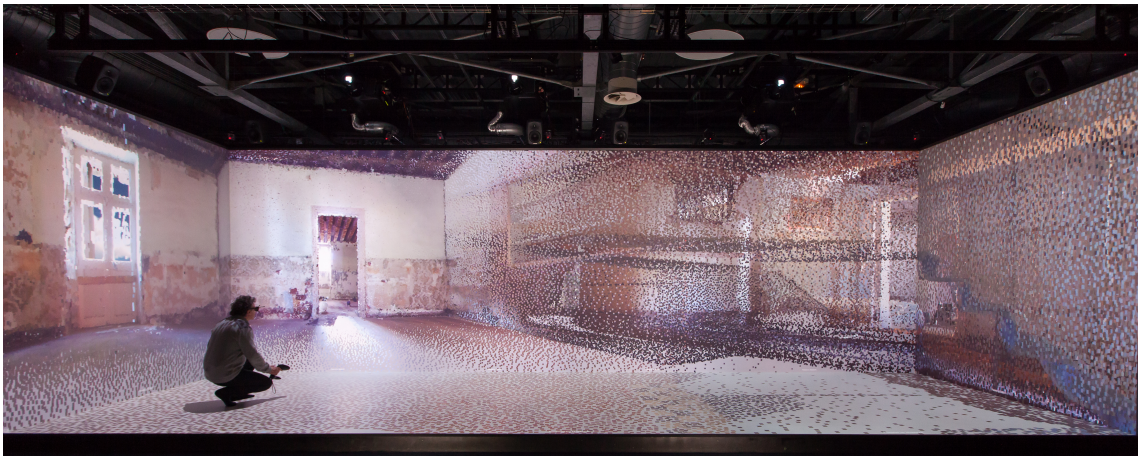


Figure 17. Immersive view of the point cloud of the Real Tennis building.

7.5.3. Making virtual archeology great again (without scientific compromise)

Participants: Ronan Gaugne, Valérie Gouranton [contact].

In the past two decades or so, digital tools have been slowly integrated as part of the archaeological process of information acquisition, analysis, and dissemination. We are now entering a new era, adding the missing piece to the puzzle in order to complete this digital revolution and take archaeology one step further into virtual reality (VR). The main focus of this work is the methodology of digital archaeology that fully integrates virtual reality, from beta testing to interdisciplinary teamwork. After data acquisition and processing necessary to construct the 3D model, we explore the analysis that can be conducted during and after the making or creation of the 3D environment and the dissemination of knowledge. We explain the relevance of this methodology through the case study on the intendant's palace, an 18th century archaeological site in Quebec City, Canada (Figure 18 left). With this experience, we believe that VR can prompt new questions that would never have occurred otherwise and can provide technical advantages in terms of gathering data in the same virtual space (Figure 18 right). We conclude that multidisciplinary input in archaeological research is once again proven essential in this new, inclusive and vast digital structure of possibilities [29].
This work was done in collaboration with UMR CREAAH, Inrap, France and Univ. Laval, Canada.

7.5.4. Evaluation of a Mixed Reality based Method for Archaeological Excavation Support

Participants: Ronan Gaugne [contact], Quentin Petit, Valérie Gouranton.

Figure 18. Left, model of the XVIIth century Palais de l'Intendant. Right, study of the reconstitution of the Palais de L'intendant and its neighborhood inside Immersia

In the context of archaeology, most of the time, micro-excavation for the study of furniture (metal, ceramics...) or archaeological context (incineration, bulk sampling) is performed without complete knowledge of the internal content, with the risk of damaging nested artifacts during the process. The use of medical imaging coupled with digital 3D technologies, has led to significant breakthroughs by allowing to refine the reading of complex artifacts. However, archaeologists may have difficulties in constructing a mental image in 3 dimensions from the axial and longitudinal sections obtained during medical imaging, and in the same way to visualize and manipulate a complex 3D object on screen, and an inability to simultaneously manipulate and analyze a 3D image, and a real object. Thereby, if digital technologies allow a 3D visualization (stereoscopic screen, VR headset ...), they are not without limiting the natural, intuitive and direct 3D perception of the archaeologist on the material or context being studied. We therefore propose a visualization system based on optical see-through augmented reality that associates real visualization of archaeological material with data from medical imaging [18] (see Figure 19). This represents a relevant approach for composite or corroded objects or contexts associating several objects such as cremations. The results presented in the paper identify adequate visualization modalities to allow archaeologist to estimate, with an acceptable error, the position of an internal element in a particular archaeological material, an Iron-Age cremation block inside a urn.
This work was done in collaboration with Inrap, France and AIST (National Institute of Advanced Industrial Science and Technology), Japan.

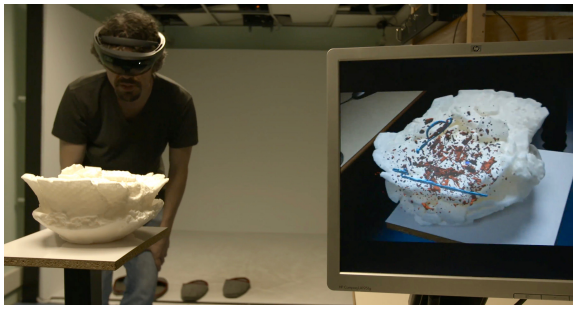


Figure 19. Evaluation of the mixed reality system

LACODAM Project-Team

7. New Results

7.1. Introduction

In this section, we organize the bulk of our contributions this year along two of our research axes, namely Pattern Mining and Decision Support. Some other contributions lie within the domain of machine learning.

7.1.1. Pattern Mining

In the domain of pattern mining we can categorize our contributions along the following lines:

- *Efficient Pattern Mining (Sections 7.2-7.4).* In [9], we propose a method to accelerate itemset sampling on FPGAs, whereas [18] proposes SSDPS, an efficient algorithm to mine discriminant patterns in two-class datasets, common in genetic data. Finally [11] presents a succinct data structure that represents concisely a cube of skypatterns.
- *Semantics of Pattern Mining (Sections 7.5-7.6).*[14] discusses the ambiguity of the semantics of pattern mining with absent events (negated statements). Likewise [8] shows formal properties of admissible generalizations in pattern mining and machine learning.

7.1.2. Decision Support

In regards to the axis of decision support, our contributions can be organized in two categories: forecasting & prediction, and modelisation.

- *Forecasting & Prediction (Sections 7.7-7.9).* In [10], we propose solutions to automate the task of capacity planning in the context of a large data network as the one available at Orange. [17] applies machine learning techniques for estrus detection in diary farms. [21] proposes a machine learning architecture in multi-sensor environments for earthquake early warning.
- *Modelling (Section 7.10).* In [5] we present a modeling approach for the nutritional requirements of lactating sows.
- *Data Exploration (Section 7.11).*[6] proposes a formal framework for the exploration of care trajectories in medical databases.

7.1.3. Others

- *Machine Learning (Section 7.12-7.14).*[7], [16] proposes novel methods to optimize the F-measure in ML, and to improve the task of domain adaptation by source selection. [19] proposes the use of GANs to make time series classification more interpretable.

7.2. Accelerating Itemset Sampling using Satisfiability Constraints on FPGA

Finding recurrent patterns within a data stream is important for fields as diverse as cybersecurity or e-commerce. This requires to use pattern mining techniques. However, pattern mining suffers from two issues. The first one, known as “pattern explosion”, comes from the large combinatorial space explored and is the result of too many patterns output to be analyzed. Recent techniques called output space sampling solve this problem by outputting only a sampled set of all the results, with a target size provided by the user. The second issue is that most algorithms are designed to operate on static datasets or low throughput streams. In [9], we propose a contribution to tackle both issues, by designing an FPGA accelerator for pattern mining with output space sampling. We show that our accelerator can outperform a state-of-the-art implementation on a server class CPU using a modest FPGA product.

7.3. Statistically Significant Discriminative Patterns Searching

In [18], we propose a novel algorithm, named SSDPS, to discover patterns in two-class datasets. The SSDPS algorithm owes its efficiency to an original enumeration strategy of the patterns, which allows to exploit some degrees of anti-monotonicity on the measures of discriminance and statistical significance. Experimental results demonstrate that the performance of the SSDPS algorithm is better than others. In addition, the number of generated patterns is much less than the number of the other algorithms. Experiment on real data also shows that SSDPS effciently detects multiple SNPs combinations in genetic data.

7.4. Compressing and Querying Skypattern Cubes

Skypatterns are important since they enable to take into account user preference through Pareto-dominance. Given a set of measures, a skypattern query finds the patterns that are not dominated by others. In practice, different users may be interested in different measures, and issue queries on any subset of measures (a.k.a subspace). This issue was recently addressed by introducing the concept of skypattern cubes. However, such a structure presents high redundancy and is not well adapted for updating operations like adding or removing measures, due to the high costs of subspace computations in retrieving skypatterns. In [11], we propose a new structure called Compressed Skypattern Cube (abbreviated CSKYC), which concisely represents a skypattern cube, and gives an efficient algorithm to compute it. We thoroughly explore its properties and provide an efficient query processing algorithm. Experimental results show that our proposal allows to construct and to query a CSKYC very efficiently.

7.5. Semantics of Negative Sequential Patterns

In the field of pattern mining, a negative sequential pattern expresses behavior by a sequence of present and absent events. In [14], we shed light on the ambiguity of this notation and identify eight possible semantics with the relation of inclusion of a motif in a sequence. These semantics are illustrated and we are studying them formally. We thus propose dominance and equivalence relationships between these semantics, and we highlight new properties of anti-monotony. These results could be used to develop new efficient algorithms for mining frequent negative sequential patterns.

7.6. Admissible Generalizations of Examples as Rules

Rule learning is a data analysis task that consists in extracting rules that generalize examples. This is achieved by a plethora of algorithms. Some generalizations make more sense for the data scientists, called here admissible generalizations. The purpose of our work in [8] is to show formal properties of admissible generalizations. A formalization for generalization of examples is proposed allowing the expression of rule admissibility. Some admissible generalizations are captured by preclosure and capping operators. Also, we are interested in selecting supersets of examples that induce such operators. We then define classes of selection functions. This formalization is more particularly developed for examples with numerical attributes. Classes of such functions are associated with notions of generalization and they are used to comment some results of the CN2 algorithm [22].

7.7. Towards a Framework for Seasonal Time Series Forecasting Using Clustering

Seasonal behaviours are widely encountered in various applications. For instance, requests on web servers are highly influenced by our daily activities. Seasonal forecasting consists in forecasting the whole next season for a given seasonal time series. It may help a service provider to provision correctly the potentially required resources, avoiding critical situations of over or under provision. In [10], we propose a generic framework to make seasonal time series forecasting. The framework combines machine learning techniques 1) to identify the typical seasons and 2) to forecast the likelihood of having a season type in one season ahead. We study this framework by comparing the mean squared errors of forecasts for various settings and various datasets. The best setting is then compared to state-of-the-art time series forecasting methods. We show that it is competitive with then.

7.8. Towards Sustainable Dairy Management - A Machine Learning Enhanced Method for Estrus Detection

Our research tackles the challenge of milk production resource use efficiency in dairy farms with machine learning methods. Reproduction is a key factor for dairy farm performance since cows milk production begin with the birth of a calf. Therefore, detecting estrus, the only period when the cow is susceptible to pregnancy, is crucial for farm efficiency. Our goal is to enhance estrus detection (performance, interpretability), especially on the currently undetected silent estrus (35% of total estrus), and allow farmers to rely on automatic estrus detection solutions based on affordable data (activity, temperature). In [17] we first propose a novel approach with real-world data analysis to address both behavioral and silent estrus detection through machine learning methods. Second, we present LCE, a local cascade based algorithm that significantly outperforms a typical commercial solution for estrus detection, driven by its ability to detect silent estrus. Then, our study reveals the pivotal role of activity sensors deployment in estrus detection. Finally, we propose an approach relying on global and local (behavioral versus silent) algorithm interpretability (SHAP) to reduce the mistrust in estrus detection solutions.

7.9. A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning

Our research [21] aims to improve the accuracy of Earthquake Early Warning (EEW) systems by means of machine learning. EEW systems are designed to detect and characterize medium and large earthquakes before their damaging effects reach a certain location. Traditional EEW methods based on seismometers fail to accurately identify large earthquakes due to their sensitivity to the ground motion velocity. The recently introduced high-precision GPS stations, on the other hand, are ineffective to identify medium earthquakes due to its propensity to produce noisy data. In addition, GPS stations and seismometers may be deployed in large numbers across different locations and may produce a significant volume of data consequently, affecting the response time and the robust-ness of EEW systems. In practice, EEW can be seen as a typical classification problem in the machine learning field: multi-sensor data are given in input, and earthquake severity is the classification result. In this paper, we introduce the Distributed Multi-Sensor Earthquake Early Warning (DMSEEW) system, a novel machine learning-based approach that combines data from both types of sensors (GPS stations and seismometers) to detect medium and large earthquakes. DMSEEW is based on a new stacking ensemble method which has been evaluated on a real-world dataset validated with geoscientists. The system builds on a geographically distributed infrastructure, ensuring an efficient computation in terms of response time and robustness to partial infrastructure failures. Our experiments show that DMSEEW is more accurate than the traditional seismometer-only approach and the combined-sensors (GPS and seis-mometers) approach that adopts the rule of relative strength.

7.10. Dynamic Modeling of Nutrient Use and Individual Requirements of Lactating Sows

Nutrient requirements of sows during lactation are related mainly to their milk yield and feed intake, and vary greatly among individuals. In practice, nutrient requirements are generally determined at the population level based on average performance. The objective of the present modeling approach was to explore the variability in nutrient requirements among sows by combining current knowledge about nutrient use with on-farm data available on sows at farrowing [parity, BW, backfat thickness (BT)] and their individual performance (litter size, litter average daily gain, daily sow feed intake) to estimate nutrient requirements. The approach was tested on a database of 1,450 lactations from 2 farms. The effects of farm (A, B), week of lactation (W1: week 1, W2: week 2, W3+: week 3 and beyond), and parity (P1: 1, P2: 2, P3+: 3 and beyond) on sow performance and their nutrient requirements were evaluated. The mean daily ME requirement was strongly correlated with litter growth ($R2 = 0.95$; $P < 0.001$) and varied slightly according to sow BW, which influenced the maintenance cost. The mean daily standardized ileal digestible (SID) lysine requirement was influenced by farm, week of lactation, and parity. Variability in SID lysine requirement per kg feed was related mainly to feed intake

(R2 = 0.51; P < 0.001) and, to a smaller extent, litter growth (R2 = 0.27; P < 0.001). It was lowest in W1 (7.0 g/kg), greatest in W2 (7.9 g/kg), and intermediate in W3+ (7.5 g/kg; P < 0.001) because milk production increased faster than feed intake capacity did. It was lower for P3+ (6.7 g/kg) and P2 sows (7.3 g/kg) than P1 sows (8.3 g/kg) due to the greater feed intake of multiparous sows. The SID lysine requirement per kg of feed was met for 80% of sows when supplies were 112 and 120% of the mean population requirement on farm A and B, respectively, indicating higher variability in requirements on farm B. Other amino acid and mineral requirements were influenced in the same way as SID lysine. In [5], we present a modeling approach that allows us to capture individual variability in the performance of sows and litters according to farm, stage of lactation, and parity. It is an initial step in the development of new types of models able to process historical farm data (e.g., for ex post assessment of nutrient requirements) and real-time data (e.g., to control precision feeding).

7.11. Temporal Models of Care Sequences for the Exploration of Medico-administrative Data

Pharmaco-epidemiology with medico-administrative databases enables the study of the impact of health products in real-life settings. These studies require to manipulate the raw data and the care trajectories, in order to identify pieces of data that may witness the medical information that is looked for. The manipulation can be seen as a querying process in which a query is a description of a medical pattern (e.g. occurrence of illness) with the available raw features from care trajectories (e.g. occurrence of medical procedures, drug deliveries, etc.). The more expressive is the querying process, the more accurate is the medical pattern search. The temporal dimension of care trajectories is a potential information that may improve the description of medical patterns. The objective of this work [6] is to propose a formal framework that would design a well-founded tool for querying care trajectories with temporal medical patterns. In this preliminary work, we present the problematic and we introduce a use case that illustrates the comparison of several querying formalisms.

7.12. Improving Domain Adaptation By Source Selection

Domain adaptation consists in learning from a source data distribution a model that will be used on a different target data distribution. The domain adaptation procedure is usually unsuccessful if the source domain is too different from the target one. In [16], we study domain adaptation for image classification with deep learning in the context of multiple available source domains. This work proposes a multi-source domain adaptation method that selects and weights the sources based on inter-domain distances. We provide encouraging results on both classical benchmarks and a new real world application with 21 domains.

7.13. From Cost-Sensitive Classification to Tight F-measure Bounds

The F-measure is a classification performance measure, especially suited when dealing with imbalanced datasets, which provides a compromise between the precision and the recall of a classifier. As this measure is non-convex and non-linear, it is often indirectly optimized using cost-sensitive learning (that affects different costs to false positives and false negatives). In [7], we derive theoretical guarantees that give tight bounds on the best F-measure that can be obtained from cost-sensitive learning. We also give an original geometric interpretation of the bounds that serves as an inspiration for CONE, a new algorithm to optimize for the F-measure. Using 10 datasets exhibiting varied class imbalance, we illustrate that our bounds are much tighter than previous work and show that CONE learns models with either superior F-measures than existing methods or comparable but in fewer iterations.

7.14. Time Series Classification Based on Interpretable Shapelets

[19] proposes a new architecture, called AI⟷PR-CNN, composed of generative adversarial neural networks (GANs), which addresses the problem of the lack of interpretability of the existing methods for time series classification. Our network has two components: a classifier and a discriminator. The classifier is a CNN, it serves to classify series. Convolutions are discriminant patterns learned from the data that allow for a more

discriminating representation of time series (similar to a shapelet). To be able to explain the decision of the classifier, we would like to impose that the convolutions used are real "shapelets", that is to say that they are close to real sub-series present in the training set. This constraint is implemented by a GAN whose purpose will determine how much the weight matrices classifier convolutions are close to subset of the training set.

LINKMEDIA Project-Team

7. New Results

7.1. Extracting and Representing Information

7.1.1. Text Mining in the Clinical Domain

Participants: Clément Dalloux, Vincent Claveau.

Clinical records cannot be shared, which is a real hurdle to develop and compare information extraction techniques. In the framework of the BigClin Project we have developed annotated corpora, that share the same linguistic properties than records, but can be freely distributed for research purposes. Several corpora and several types of annotation were proposed for French, Portuguese and English. They are made freely available for research purposes and are described in [27], [25]. These corpora will foster reproducible research on clinical text mining.

Thanks to these datasets, we have organized the DeFT text-mining competition in 2019. Several NLP techniques and tools have been developed within the project in order to identify relevant medical or linguistic information [30], [26]. They are all chiefly based on machine learning approaches, and for most of them, more specifically, on deep learning. For instance, we have developed a new Part-of-Speech tagger and lemmatizer for French, especially suited to handle medical texts; it is freely available as a web-service at https://allgo.inria.fr . The identification of negation and uncertainty is important to precisely understand the clinical texts. Thus, we have continued our work on neural techniques to find the negation/uncertainty cues and their scope (part of sentence concerned by the negation or uncertainty). It achieves state-of-the-art results on English, and is pioneer work for French and Portuguese for which it sets a new standard [4], [21]; it is available at https://allgo.inria.fr . Other achievements in text-mining include: numerical value extraction (finding concepts that are measured, such as lab results, numerical expressions, their units) in French, English and Portuguese, the identification of gender, age, outcome and admission reasons in French clinical texts, ...

7.1.2. Embedding in hyperbolic spaces

Participants: François Torregrossa, Vincent Claveau, Guillaume Gravier.

During this year, we have studied non-Euclidean spaces into which one can embed data (for instance, words). We have developed the HierarX tool which projects multiple datasources into hyperbolic manifolds: Lorentz or Poincaré. From similarities between word pairs or continuous word representations in high dimensional spaces, HierarX is able to embed knowledge in hyperbolic geometries with small dimensionality. Those shape information into continuous hierarchies. The source code is available on the Inria's GitLab.

7.1.3. Aggregation and embedding for group membership verification

Participants: Marzieh Gheisari Khorasgani, Teddy Furon, Laurent Amsaleg.

This paper proposes a group membership verification protocol preventing the curious but honest server from reconstructing the enrolled signatures and inferring the identity of querying clients [24]. The protocol quantizes the signatures into discrete embeddings, making reconstruction difficult. It also aggregates multiple embeddings into representative values, impeding identification. Theoretical and experimental results show the trade-off between the security and error rates.

7.1.4. Group Membership Verification with Privacy: Sparse or Dense?

Participants: Marzieh Gheisari Khorasgani, Teddy Furon, Laurent Amsaleg.

Group membership verification checks if a biometric trait corresponds to one member of a group without revealing the identity of that member. Recent contributions provide privacy for group membership protocols through the joint use of two mechanisms: quantizing templates into discrete embeddings, and aggregating several templates into one group representation. However, this scheme has one drawback: the data structure representing the group has a limited size and cannot recognize noisy query when many templates are aggregated. Moreover, the sparsity of the embeddings seemingly plays a crucial role on the performance verification. This contribution proposes a mathematical model for group membership verification allowing to reveal the impact of sparsity on both security, compactness, and verification performances [23]. This models bridges the gap towards a Bloom filter robust to noisy queries. It shows that a dense solution is more competitive unless the queries are almost noiseless.

7.1.5. Privacy Preserving Group Membership Verification and Identification

Participants: Marzieh Gheisari Khorasgani, Teddy Furon, Laurent Amsaleg.

When convoking privacy, group membership verification checks if a biometric trait corresponds to one member of a group without revealing the identity of that member. Similarly, group membership identification states which group the individual belongs to, without knowing his/her identity. A recent contribution provides privacy and security for group membership protocols through the joint use of two mechanisms: quantizing biometric templates into discrete embeddings, and aggregating several templates into one group representation. This paper significantly improves that contribution because it jointly learns how to embed and aggregate instead of imposing fixed and hard coded rules [10]. This is demonstrated by exposing the mathematical underpinnings of the learning stage before showing the improvements through an extensive series of experiments targeting face recognition. Overall, experiments show that learning yields an excellent trade-off between security/privacy and the verification/identification performances.

7.1.6. Intrinsic Dimensionality Estimation within Tight Localities

Participants: Laurent Amsaleg, Oussama Chelly [Microsoft Germany], Michael Houle [National Institute of Informatics, Japan], Ken-Ichi Kawarabayashi [National Institute of Informatics, Japan], Miloš Radovanović [Univ. Novi Sad, Serbia], Weeris Treeratanajaru [Chulalongkorn University, Thailand].

Accurate estimation of Intrinsic Dimensionality (ID) is of crucial importance in many data mining and machine learning tasks, including dimensionality reduction, outlier detection, similarity search and subspace clustering. However, since their convergence generally requires sample sizes (that is, neighborhood sizes) on the order of hundreds of points, existing ID estimation methods may have only limited usefulness for applications in which the data consists of many natural groups of small size. In this paper, we propose a local ID estimation strategy stable even for 'tight' localities consisting of as few as 20 sample points [31]. The estimator applies MLE techniques over all available pairwise distances among the members of the sample, based on a recent extreme-value-theoretic model of intrinsic dimensionality, the Local Intrinsic Dimension (LID). Our experimental results show that our proposed estimation technique can achieve notably smaller variance, while maintaining comparable levels of bias, at much smaller sample sizes than state-of-the-art estimators.

7.1.7. Selective Biogeography-Based Optimizer Considering Resource Allocation for Large-Scale Global Optimization

Participants: Meiji Cui [Tongji University, China], Li Li [Tongji University, China], Miaojing Shi.

Biogeography-based optimization (BBO), a recent proposed meta-heuristic algorithm, has been successfully applied to many optimization problems due to its simplicity and efficiency. However, BBO is sensitive to the curse of dimensionality; its performance degrades rapidly as the dimensionality of the search space increases. In [3], a selective migration operator is proposed to scale up the performance of BBO and we name it selective BBO (SBBO). The differential migration operator is selected heuristically to explore the global area as far as possible whist the normal distributed migration operator is chosen to exploit the local area. By the means of heuristic selection, an appropriate migration operator can be used to search the global optimum efficiently. Moreover, the strategy of cooperative co-evolution (CC) is adopted to solve large-scale global optimization problems (LSOPs). To deal with subgroup imbalance contribution to the whole solution in the context of

CC, a more efficient computing resource allocation is proposed. Extensive experiments are conducted on the CEC 2010 benchmark suite for large-scale global optimization, and the results show the effectiveness and efficiency of SBBO compared with BBO variants and other representative algorithms for LSOPs. Also, the results confirm that the proposed computing resource allocation is vital to the large-scale optimization within the limited computation budget.

7.1.8. Friend recommendation for cross marketing in online brand community based on intelligent attention allocation link prediction algorithm

Participants: Shugang Li [Shanghai University, China], Xuewei Song [Shanghai University, China], Hanyu Lu [Shanghai University, China], Linyi Zeng [Shanghai University, Industrial and Commercial Bank of China, China], Miaojing Shi, Fang Liu [Shanghai University, China].

Circle structure of online brand communities allows companies to conduct cross-marketing activities by the influence of friends in different circles and build strong and lasting relationships with customers. However, existing works on the friend recommendation in social network do not consider establishing friendships between users in different circles, which has the problems of network sparsity, neither do they study the adaptive generation of appropriate link prediction algorithms for different circle features. In order to fill the gaps in previous works, the intelligent attention allocation link prediction algorithm is proposed to adaptively build attention allocation index (AAI) according to the sparseness of the network and predict the possible friendships between users in different circles. The AAI reflects the amount of attention allocated to the user pair by their common friend in the triadic closure structure, which is decided by the friend count of the common friend. Specifically, for the purpose of overcoming the problem of network sparsity, the AAIs of both the direct common friends and indirect ones are developed. Next, the decision tree (DT) method is constructed to adaptively select the suitable AAIs for the circle structure based on the density of common friends and the dispersion level of common friends' attention. In addition, for the sake of further improving the accuracy of the selected AAI, its complementary AAIs are identified with support vector machine model according to their similarity in value, direction, and ranking. Finally, the mutually complementary indices are combined into a composite one to comprehensively portray the attention distribution of common friends of users in different circles and predict their possible friendships for cross-marketing activities. Experimental results on Twitter and Google+ show that the model has highly reliable prediction performance [5].

7.1.9. Revisiting the medial axis for planar shape decomposition

Participants: Nikos Papanelopoulos [NTUA, Greece], Yannis Avrithis, Stefanos Kollias [U. of Lincoln, UK].

We present a simple computational model for planar shape decomposition that naturally captures most of the rules and salience measures suggested by psychophysical studies, including the minima and short-cut rules, convexity, and symmetry. It is based on a medial axis representation in ways that have not been explored before and sheds more light into the connection between existing rules like minima and convexity. In particular, vertices of the exterior medial axis directly provide the position and extent of negative minima of curvature, while a traversal of the interior medial axis directly provides a small set of candidate endpoints for part-cuts. The final selection follows a prioritized processing of candidate part-cuts according to a local convexity rule that can incorporate arbitrary salience measures. Neither global optimization nor differentiation is involved. We provide qualitative and quantitative evaluation and comparisons on ground-truth data from psycho-physical experiments. With our single computational model, we outperform even an ensemble method on several other competing models [6].

7.1.10. Graph-based Particular Object Discovery

Participants: Oriane Siméoni, Ahmet Iscen [Univ. Prague], Giorgos Tolias [Univ. Prague], Yannis Avrithis, Ondra Chum [Univ. Prague].

Severe background clutter is challenging in many computer vision tasks, including large-scale image retrieval. Global descriptors, that are popular due to their memory and search efficiency, are especially prone to corruption by such a clutter. Eliminating the impact of the clutter on the image descriptor increases the chance of retrieving relevant images and prevents topic drift due to actually retrieving the clutter in the case of query expansion. In this work, we propose a novel salient region detection method. It captures, in an unsupervised manner, patterns that are both discriminative and common in the dataset. Saliency is based on a centrality measure of a nearest neighbor graph constructed from regional CNN representations of dataset images. The proposed method exploits recent CNN architectures trained for object retrieval to construct the image representation from the salient regions. We improve particular object retrieval on challenging datasets containing small objects [7].

7.1.11. Label Propagation for Deep Semi-supervised Learning

Participants: Ahmet Iscen [Univ. Prague], Giorgos Tolias [Univ. Prague], Yannis Avrithis, Ondra Chum [Univ. Prague].

Semi-supervised learning is becoming increasingly important because it can combine data carefully labeled by humans with abundant unlabeled data to train deep neural networks. Classic methods on semi-supervised learning that have focused on transductive learning have not been fully exploited in the inductive framework followed by modern deep learning. The same holds for the manifold assumption—that similar examples should get the same prediction. In this work, we employ a transductive label propagation method that is based on the manifold assumption to make predictions on the entire dataset and use these predictions to generate pseudo-labels for the unlabeled data and train a deep neural network. At the core of the transductive method lies a nearest neighbor graph of the dataset that we create based on the embeddings of the same network. Therefore our learning process iterates between these two steps. We improve performance on several datasets especially in the few labels regime and show that our work is complementary to current state of the art [12], [38].

7.1.12. Dense Classification and Implanting for Few-Shot Learning

Participants: Yann Lifchitz, Yannis Avrithis, Sylvaine Picard [SAFRAN Group], Andrei Bursuc [Valéo].

Few-shot learning for deep neural networks is a highly challenging and key problem in many computer vision tasks. In this context, we are targeting knowledge transfer from a set with abundant data to other sets with few available examples. We propose in [14], [40] two simple and effective solutions: (i) dense classification over feature maps, which for the first time studies local activations in the domain of few- shot learning, and (ii) implanting, that is, attaching new neurons to a previously trained network to learn new, task-specific features. Implanting enables training of multiple layers in the few-shot regime, departing from most related methods derived from metric learning that train only the final layer. Both contributions show consistent gains when used individually or jointly and we report state of the art performance on few-shot classification on miniImageNet.

7.1.13. Point in, Box out: Beyond Counting Persons in Crowds

Participants: Yuting Liu [Sichuan University, China], Miaojing Shi, Qijun Zhao [Sichuan University, China], Xiaofang Wang [RAINBOW Team, IRISA].

Modern crowd counting methods usually employ deep neural networks (DNN) to estimate crowd counts via density regression. Despite their significant improvements, the regression-based methods are incapable of providing the detection of individuals in crowds. The detection-based methods, on the other hand, have not been largely explored in recent trends of crowd counting due to the needs for ex- pensive bounding box annotations. In this work, we instead propose a new deep detection network with only point supervision required [15]. It can simultaneously detect the size and location of human heads and count them in crowds. We first mine useful person size information from point-level annotations and initialize the pseudo ground truth bounding boxes. An online updating scheme is introduced to refine the pseudo ground truth during training; while a locally- constrained regression loss is designed to provide additional constraints on the size of the predicted boxes in a local neighborhood. In the end, we propose a curriculum learning strategy to train the network from images of relatively accurate and easy pseudo ground truth first. Extensive experiments are conducted in both detection and counting tasks on several standard benchmarks, e.g. ShanghaiTech, UCF CC

50, WiderFace, and TRANCOS datasets, and the results show the superiority of our method over the state-of-the-art.

7.1.14. Revisiting Perspective Information for Efficient Crowd Counting

Participants: Miaojing Shi, Zhaohui Yang [Peking University, China], Chao Xu [Peking University, China], Qijun Chen [Tongji University, China].

Crowd counting is the task of estimating people numbers in crowd images. Modern crowd counting methods employ deep neural networks to estimate crowd counts via crowd density regressions. A major challenge of this task lies in the perspective distortion, which results in drastic person scale change in an image. Density regression on the small person area is in general very hard. In this work, we propose a perspective-aware convolutional neural network (PACNN) for efficient crowd counting, which integrates the perspective information into density regression to provide additional knowledge of the person scale change in an image [18]. Ground truth perspective maps are firstly generated for training; PACNN is then specifically designed to predict multi-scale perspective maps, and encode them as perspective-aware weighting layers in the network to adaptively combine the outputs of multi-scale density maps. The weights are learned at every pixel of the maps such that the final density combination is robust to the perspective distortion. We conduct extensive experiments on the ShanghaiTech, WorldExpo'10, UCF CC 50, and UCSD datasets, and demonstrate the effectiveness and efficiency of PACNN over the state-of-the-art.

7.1.15. Local Features and Visual Words Emerge in Activations

Participants: Oriane Siméoni, Yannis Avrithis, Ondra Chum [Univ. Prague].

We propose a novel method of deep spatial matching (DSM) for image retrieval [19], [41]. Initial ranking is based on image descriptors extracted from convolutional neural network activations by global pooling, as in recent state-of- the-art work. However, the same sparse 3D activation tensor is also approximated by a collection of local features. These local features are then robustly matched to approximate the optimal alignment of the tensors. This hap- pens without any network modification, additional layers or training. No local feature detection happens on the original image. No local feature descriptors and no visual vocabulary are needed throughout the whole process. We experimentally show that the proposed method achieves the state-of-the-art performance on standard benchmarks across different network architectures and different global pooling methods. The highest gain in performance is achieved when diffusion on the nearest-neighbor graph of global descriptors is initiated from spatially verified images.

7.1.16. Combining convolutional side-outputs for road image segmentation

Participants: Raquel Almeida, Simon Malinowski, Ewa Kijak, Silvio Guimaraes [PUC Minas].

Image segmentation consists in creating partitions within an image into meaningful areas and objects. It can be used in scene understanding and recognition, in fields like biology, medicine, robotics, satellite imaging, amongst others. In this work [17], we take advantage of the learned model in a deep architecture, by extracting side-outputs at different layers of the network for the task of image segmentation. We study the impact of the amount of side-outputs and evaluate strategies to combine them. A post-processing filtering based on mathematical morphology idempotent functions is also used in order to remove some undesirable noises. Experiments were performed on the publicly available KITTI Road Dataset for image segmentation. Our comparison shows that the use of multiples side outputs can increase the overall performance of the network, making it easier to train and more stable when compared with a single output in the end of the network. Also, for a small number of training epochs (500), we achieved a competitive performance when compared to the best algorithm in KITTI Evaluation Server.

7.1.17. BRIEF-based mid-level representations for time series classification

Participants: Raquel Almeida, Simon Malinowski, Silvio Guimaraes [PUC Minas].

Time series classification has been widely explored over the last years. Amongst the best approaches for that task, many are based on the Bag-of-Words framework, in which time series are transformed into a histogram of word occurrences. These words represent quantized features that are extracted beforehand. In this work [20], we aim to evaluate the use of accurate mid-level representation called BossaNova in order to enhance the Bag-of-Words representation and to propose a new binary time series descriptor, called BRIEF-based descriptor. More precisely, this kind of representation enables to reduce the loss induced by feature quantization. Experiments show that this representation in conjunction to BRIEF-based descriptor is statistically equivalent to traditional Bag-of-Words, in terms time series classification accuracy, being about 4 times faster. Furthermore, it is very competitive when compared to the state-of-the-art.

7.1.18. Toward a Framework for Seasonal Time Series Forecasting Using Clustering

Participants: Simon Malinowski, Thomas Guyet [LACODAM Team], Colin Leverger [LACODAM Team], Alexandre Termier [LACODAM Team].

Seasonal behaviours are widely encountered in various applications. For instance, requests on web servers are highly influenced by our daily activities. Seasonal forecasting consists in forecasting the whole next season for a given seasonal time series. It may help a service provider to provision correctly the potentially required resources, avoiding critical situations of over- or under provision. In this article, we propose a generic framework to make seasonal time series forecasting. The framework combines machine learning techniques (1) to identify the typical seasons and (2) to forecast the likelihood of having a season type in one season ahead. We study in[13] this framework by comparing the mean squared errors of forecasts for various settings and various datasets. The best setting is then compared to state-of-the-art time series forecasting methods. We show that it is competitive with them.

7.1.19. Smooth Adversarial Examples

Participants: Hanwei Zhang, Yannis Avrithis, Teddy Furon, Laurent Amsaleg.

This paper investigates the visual quality of the adversarial examples. Recent papers propose to smooth the perturbations to get rid of high frequency artefacts. In this work, smoothing has a different meaning as it perceptually shapes the perturbation according to the visual content of the image to be attacked [44]. The perturbation becomes locally smooth on the flat areas of the input image, but it may be noisy on its textured areas and sharp across its edges. This operation relies on Laplacian smoothing, well-known in graph signal processing, which we integrate in the attack pipeline. We benchmark several attacks with and without smoothing under a white-box scenario and evaluate their transferability. Despite the additional constraint of smoothness, our attack has the same probability of success at lower distortion.

7.1.20. Walking on the Edge: Fast, Low-Distortion Adversarial Examples

Participants: Hanwei Zhang, Yannis Avrithis, Teddy Furon, Laurent Amsaleg.

Adversarial examples of deep neural networks are receiving ever increasing attention because they help in understanding and reducing the sensitivity to their input. This is natural given the increasing applications of deep neural networks in our everyday lives. When white-box attacks are almost always successful, it is typically only the distortion of the perturbations that matters in their evaluation. In this work [45], we argue that speed is important as well, especially when considering that fast attacks are required by adversarial training. Given more time, iterative methods can always find better solutions. We investigate this speed-distortion trade-off in some depth and introduce a new attack called boundary projection (BP) that improves upon existing methods by a large margin. Our key idea is that the classification boundary is a manifold in the image space: we therefore quickly reach the boundary and then optimize distortion on this manifold.

7.1.21. Accessing watermarking information: Error exponents in the noisy case

Participant: Teddy Furon.

The study of the error exponents of zero-bit watermarking is addressed in the article by Comesana, Merhav, and Barni, under the assumption that the detector relies solely on second order joint empirical statistics of the received signal and the watermark. This restriction leads to the well-known dual hypercone detector, whose score function is the absolute value of the normalized correlation. They derive the false negative error exponent and the optimum embedding rule. However, they only focus on high SNR regime, i.e. the noiseless scenario. This work extends this theoretical study to the noisy scenario. It introduces a new definition of watermarking robustness based on the false negative error exponent, derives this quantity for the dual hypercone detector, and shows that its performances is almost equal to Costa's lower bound [22].

7.1.22. Detecting fake news and image forgeries

Participants: Cédric Maigrot, Vincent Claveau, Ewa Kijak.

Social networks make it possible to share information rapidly and massively. Yet, one of their major drawback comes from the absence of verification of the piece of information, especially with viral messages. Based on the work already presented in the previous years, C. Maigrot defended his thesis on the detection of image forgeries, classification of reinformation websites, and on the late fusion of models based on the text, image and source analysis [1]. This work was also given a large visibility thanks to numerous interviews in Press and TV (see the dedicated section about popularization).

7.1.23. Learning Interpretable Shapelets for Time Series Classification through Adversarial Regularization

Times series classification can be successfully tackled by jointly learning a shapelet-based representation of the series in the dataset and classifying the series according to this representation. However, although the learned shapelets are discriminative, they are not always similar to pieces of a real series in the dataset. This makes it difficult to interpret the decision, i.e. difficult to analyze if there are particular behaviors in a series that triggered the decision. In this work [29], we make use of a simple convolutional network to tackle the time series classification task and we introduce an adversarial regularization to constrain the model to learn more interpretable shapelets. Our classification results on all the usual time series benchmarks are comparable with the results obtained by similar state-of-the-art algorithms but our adversarially regularized method learns shapelets that are, by design, interpretable.

7.1.24. Using Knowledge Base Semantics in Context-Aware Entity Linking

Participants: Cheikh Brahim El Vaigh, Guillaume Gravier, Pascale Sébillot.

Done as part of the IPL iCODA, in collaboration with CEDAR Inria team.

Entity linking is a core task in textual document processing, which consists in identifying the entities of a knowledge base (KB) that are mentioned in a text. Approaches in the literature consider either independent linking of individual mentions or collective linking of all mentions. Regardless of this distinction, most approaches rely on the Wikipedia encyclopedic KB in order to improve the linking quality, by exploiting its entity descriptions (web pages) or its entity interconnections (hyperlink graph of web pages). We devised a novel collective linking technique which departs from most approaches in the literature by relying on a structured RDF KB [9]. This allows exploiting the semantics of the interrelationships that candidate entities may have at disambiguation time rather than relying on raw structural approximation based on Wikipedia's hyperlink, graph. The few approaches that also use an RDF KB simply rely on the existence of a relation between the candidate entities to which mentions may be linked. Instead, we weight such relations based on the RDF KB structure and propose an efficient decoding strategy for collective linking. Experiments on standard benchmarks show significant improvement over the state of the art.

7.1.25. Neural-based lexico-syntactic relation extraction in news archives

Participants: Guillaume Gravier, Cyrielle Mallart, Pascale Sébillot.

Done as part of the IPL iCODA, in collaboration with Ouest France

Relation extraction is the task of finding and classifying the relationship between two entities in a text. We pursued work on the detection of relations between entities, seen as a binary classification problem. In the context of large-scale news archives, we argue that detection is paramount before even considering classification, where most approaches consider the two tasks jointly with a null garbage class. This does hardly allow for the detection of relations for unseen categories, which are all considered as garbage. We designed a bi-LSTM sequence neural model acting on features extracted from the surface realization, the part-of-speech tags and the dependency parse tree and compared with a state-of-the-art relation detection LSTM-based approach. Experimental evaluations rely on a dataset derived from 200k Wikipedia articles in French containing 4M linked mentions of entities: 330k pairs of entities co-occur in the same sentence, of which 1 % are actual relations according to Wikidata. Results show the benefit of our binary detection approach over previous methods and over joint detection and classification.

7.1.26. Graph Convolutional Networks for Learning with Few Clean and Many Noisy Labels

Participants: Ahmet Iscen [Google Research], Giorgos Tolias [Univ. Prague], Yannis Avrithis, Ondra Chum [Univ. Prague], Cordelia Schmid [Google Research].

In this work we consider the problem of learning a classifier from noisy labels when a few clean labeled examples are given [39]. The structure of clean and noisy data is modeled by a graph per class and Graph Convolutional Networks (GCN) are used to predict class relevance of noisy examples. For each class, the GCN is treated as a binary classifier learning to discriminate clean from noisy examples using a weighted binary cross-entropy loss function, and then the GCN-inferred "clean" probability is exploited as a relevance measure. Each noisy example is weighted by its relevance when learning a classifier for the end task. We evaluate our method on an extended version of a few-shot learning problem, where the few clean examples of novel classes are supplemented with additional noisy data. Experimental results show that our GCN-based cleaning process significantly improves the classification accuracy over not cleaning the noisy data and standard few-shot classification where only few clean examples are used. The proposed GCN-based method outperforms the transductive approach (Douze et al., 2018) that is using the same additional data without labels.

7.1.27. Rethinking deep active learning: Using unlabeled data at model training

Participants: Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, Guillaume Gravier.

Active learning typically focuses on training a model on few labeled examples alone, while unlabeled ones are only used for acquisition. In this work we depart from this setting by using both labeled and unlabeled data during model training across active learning cycles [42]. We do so by using unsupervised feature learning at the beginning of the active learning pipeline and semi-supervised learning at every active learning cycle, on all available data. The former has not been investigated before in active learning, while the study of latter in the context of deep learning is scarce and recent findings are not conclusive with respect to its benefit. Our idea is orthogonal to acquisition strategies by using more data, much like ensemble methods use more models. By systematically evaluating on a number of popular acquisition strategies and datasets, we find that the use of unlabeled data during model training brings a spectacular accuracy improvement in image classification, compared to the differences between acquisition strategies. We thus explore smaller label budgets, even one label per class.

7.1.28. Training Object Detectors from Few Weakly-Labeled and Many Unlabeled Images

Participants: Zhaohui Yang [Peking University], Miaojing Shi, Yannis Avrithis, Chao Xu [Peking University], Vittorio Ferrari [Google Research].

Weakly-supervised object detection attempts to limit the amount of supervision by dispensing the need for bounding boxes, but still assumes image-level labels on the entire training set are available. In this work, we study the problem of training an object detector from one or few clean images with image-level labels and a larger set of completely unlabeled images [43]. This is an extreme case of semi-supervised learning where the labeled data are not enough to bootstrap the learning of a classifier or detector. Our solution is to use a standard weakly-supervised pipeline to train a student model from image-level pseudo-labels generated on the unlabeled set by a teacher model, bootstrapped by region-level similarities to clean labeled images. By using

the recent pipeline of PCL and more unlabeled images, we achieve performance competitive or superior to many state of the art weakly-supervised detection solutions.

7.2. Accessing Information

7.2.1. Ontological modeling of human reading experience

Participants: Guillaume Gravier, Pascale Sébillot.

Done as part of the JPI CH READ-IT projects, in collaboration with Open University (UK) and Université Le Mans (FR)

Diaries, correspondence and authors' libraries provide important evidence into the evolution of ideas and society. Studying these phenomena is connected to understanding changes of perspective and values. Within the framework of the READ-IT project, we developed an ontological data approach modelling changes in the contents of diaries, correspondence and authors' libraries related to reading. By considering these three types of sources, we designed a conceptual data model to permit the study and increase the usability of sources containing evidence of reading experiences, highlighting common challenges and patterns related to changes to readers and to the medium of reading when confronting historical events [36], [8].

7.2.2. Integration of Exploration and Search: A Case Study of the M^3 Model

Participants: Snorri Gíslason [IT Univ. Copenhagen], Björn Þór Jónsson [IT Univ. Copenhagen], Laurent Amsaleg.

Effective support for multimedia analytics applications requires exploration and search to be integrated seamlessly into a single interaction model. Media metadata can be seen as defining a multidimensional media space, casting multimedia analytics tasks as exploration, manipulation and augmentation of that space. We present an initial case study of integrating exploration and search within this multidimensional media space [11]. We extend the M^3 model, initially proposed as a pure exploration tool, and show that it can be elegantly extended to allow searching within an exploration context and exploring within a search context. We then evaluate the suitability of relational database management systems, as representatives of today's data management technologies, for implementing the extended M^3 model. Based on our results, we finally propose some research directions for scalability of multimedia analytics.

7.2.3. Exquisitor: Breaking the Interaction Barrier for Exploration of 100 Million Images

Participants: Hanna Ragnarsdóttir [Reykjavik University], Þórhildur Þorleiksdóttir [Reykjavik University], Omar Shahbaz Khan [IT Univ. Copenhagen], Björn Þór Jónsson [IT Univ. Copenhagen], Gylfi Þór Gudmundsson [School of Computer Science, Reykjavik], Jan Zahálka [bohem.ai], Stevan Rudinac [University of Amsterdam], Laurent Amsaleg, Marcel Worring [University of Amsterdam].

We present Exquisitor, a media explorer capable of learning user preferences in real-time during interactions with the 99.2 million images of YFCC100M. Exquisitor owes its efficiency to innovations in data representation, compression, and indexing. Exquisitor can complete each interaction round, including learning preferences and presenting the most relevant results, in less than 30 ms using only a single CPU core and modest RAM. In short, Exquisitor can bring large-scale interactive learning to standard desktops and laptops, and even high-end mobile devices [16].

MIMETIC Project-Team

7. New Results

7.1. Outline

In 2019, MimeTIC has maintained his activity in motion analysis, modelling and simulation, to support the idea that these approaches are strongly coupled in a motion analysis-synthesis loop. This idea has been applied to the main application domains of MimeTIC:

- Animation, autonomous characters and Digital Storytelling,
- Fidelity of Virtual Reality,
- Motion sensing of Human Activity,
- Sports,
- Ergonomics,
- and Locomotion and Interactions between walkers.

7.2. Animation, Autonomous Characters and Digital Storytelling

MimeTIC main research path consists in associating motion analysis and synthesis to enhance the naturalness in computer animation, with applications in movie previsualisation, and autonomous virtual character control. Thus, we pushed example-based techniques in order to reach a good tradeoff between simulation efficiency and naturalness of the results. In 2019, to achieve this goal, MimeTIC continued to explore the use of perceptual studies and model-based approaches, but also began to investigate deep learning, for example to control cameras in Movie previsualization.

7.2.1. VR as a Content Creation Tool for Movie Previsualisation

Participants: Marc Christie [contact], Quentin Galvane.

This work proposes a VR authoring system which provides intuitive ways of crafting visual sequences in 3D environments, both for expert animators and expert creatives. It is designed in mind to be applied animation and film industries, but can find broader applications (eg. in multimedia content creation). Creatives in animation and film productions have forever been exploring the use of new means to prototype their visual sequences before realizing them, by relying on hand-drawn storyboards, physical mockups or more recently 3D modelling and animation tools. However these 3D tools are designed in mind for dedicated animators rather than creatives such as film directors or directors of photography and remain complex to control and master. The proposed system is designed to reflect the traditional process through (i) a storyboarding mode that enables rapid creation of annotated still images, (ii) a previsualisation mode that enables the animation of the characters, objects and cameras, and (iii) a technical mode that enables the placement and animation of complex camera rigs (such as cameras cranes) and light rigs. Our methodology strongly relies on the benefits of VR manipulations to re-think how content creation can be performed in this specific context, typically how to animate contents in space and time. As a result, the proposed system is complimentary to existing tools, and provides a seamless back-and-forth process between all stages of previsualisation. We evaluated the tool with professional users to gather experts' perspectives on the specific benefits of VR in 3D content creation [36].

7.2.2. Deep Learning Techniques for Camera Trajectories

Participant: Marc Christie [contact].

Designing a camera motion controller which places and moves virtual cameras in relation with contents in a cinematographic way is a complex and challenging task. Many cinematographic rules exist, yet practice shows there are significant stylistic variations in how these can be applied. While contributions have attempted to encode rules by hand, this work is the very first to propose an end-to-end framework that automatically learns from real and synthetic movie sequences how the camera behaves in relation with contents. Our deep-learning framework extracts cinematic features of movies through a novel feature estimator trained on synthetic data, and learns camera behaviors from those extracted features, through the design of a Recurrent Neural Network (RNN) with a Mixture of Experts (MoE) gating mechanism. This cascaded network is designed to capture important variations in camera behaviors while ensuring the generalization capacity in the learning of similar behaviors. We demonstrate the features of our framework through experiments that highlight (i) the quality of our cinematic feature extractor (ii) the capacity to learn ranges of behaviors through the gating mechanism, and (iii) the ability to analyse the camera behaviors from a given input sequence, and automatically re-apply these behaviors on new virtual contents, offering exciting new possibilities towards a deeper understanding of cinematographic style and enhanced possibilities in transferring style from real to virtual. The work is a collaboration with the Beijing Film Academy in China.

7.2.3. Efficient Visibility Computation for Camera Control

Participants: Marc Christie [contact], Ludovic Burg.

Efficient visibility computation is a prominent requirement when designing automated camera control techniques for dynamic 3D environments; computer games, interactive storytelling or 3D media applications all need to track 3D entities while ensuring their visibility and delivering a smooth cinematographic experience. Addressing this problem requires to sample a very large set of potential camera positions and estimate visibility for each of them, which in practice is intractable. In this work, we introduce a novel technique to perform efficient visibility computation and anticipate occlusions. We first propose a GPU-rendering technique to sample visibility in Toric Space coordinates – a parametric space designed for camera control. We then rely on this visibility evaluation to compute an anticipation map which predicts the future visibility of a large set of cameras over a specified number of frames. We finally design a camera motion strategy that exploits this anticipation map to maximize the visibility of entities over time. The key features of our approach are demonstrated through comparison with classical ray-casting techniques on benchmark environments, and through an integration in multiple game-like 3D environment with heavy sparse and dense occluders.

7.2.4. Analysing and Predicting Inter-Observer Gaze Congruency

Participant: Marc Christie [contact].

In trying to better understand film media, we have been recently exploring the relation between the distribution of gaze states and the features of images, with the objective of establishing correlations to understand how films manipulate users gaze (and how gaze can be manipulated be re-editing film sequences). According to the literature regarding visual saliency, observers may exhibit considerable variations in their gaze behaviors. These variations are influenced by aspects such as cultural background, age or prior experiences, but also by features in the observed images. The dispersion between the gaze of different observers looking at the same image is commonly referred as inter-observer congruency (IOC). Predicting this congruence can be of great interest when it comes to study the visual perception of an image. We introduce a new method based on deep learning techniques to predict the IOC of an image [31]. This is achieved by first extracting features from an image through a deep convolutional network. We then show that using such features to train a model with a shallow network regression technique significantly improves the precision of the prediction over existing approaches.

7.2.5. Deep Saliency Models: the Quest for the Loss Function

Participant: Marc Christie [contact].

Following our idea of understanding gaze patterns in movie watching, and predicting these gaze patterns on sequences, we have been exploring the influence of loss functions in learning the visual saliency. Indeed, numerous models in the literature present new ways to design neural networks, to arrange gaze pattern data, or to extract as much high and low-level image features as possible in order to create the best saliency representation. However, one key part of a typical deep learning model is often neglected: the choice of the loss function. In this work, we explore some of the most popular loss functions that are used in deep saliency models [49]. We demonstrate that on a fixed network architecture, modifying the loss function can significantly improve (or depreciate) the results, hence emphasizing the importance of the choice of the loss function when designing a model. We also introduce new loss functions that have never been used for saliency prediction to our knowledge. And finally, we show that a linear combination of several well-chosen loss functions leads to significant improvements in performances on different datasets as well as on a different network architecture, hence demonstrating the robustness of a combined metric.

7.2.6. Contact Preserving Shape Transfer For Rigging-Free Motion Retargeting

Participants: Franck Multon [contact], Jean Basset.

In 2018, we introduced the idea of context graph to capture the relationship between body parts surfaces and enhance the quality of the motion retargetting problem. Hence, it becomes possible to retarget the motion of a source character to a target one while preserving the topological relationship between body parts surfaces. However this approach implies to strictly satisfy distance constraints between body parts, whereas some of them could be relaxed to preserve naturalness. In 2019, we introduced a new paradigm based on transfering the shape instead of encoding the pose constraints to tackle this problem [29].

Hence, retargeting a motion from a source to a target character is an important problem in computer animation, as it allows to reuse existing rigged databases or transfer motion capture to virtual characters. Surface based pose transfer is a promising approach to avoid the trial-and-error process when controlling the joint angles. The main contribution of this work is to investigate whether shape transfer instead of pose transfer would better preserve the original contextual meaning of the source pose. To this end, we propose an optimization-based method to deform the source shape+pose using three main energy functions: similarity to the target shape, body part volume preservation, and collision management (preserve existing contacts and prevent penetrations). The results show that our method is able to retarget complex poses, including several contacts, to very different morphologies. In particular, we introduce new contacts that are linked to the change in morphology, and which would be difficult to obtain with previous works based on pose transfer that aim at distance preservation between body parts. These preliminary results are encouraging and open several perspectives, such as decreasing computation time, and better understanding how to model pose and shape constraints.

7.2.7. The Influence of Step Length to Step Frequency Ratio on the Perception of Virtual Walking Motions

Participants: Ludovic Hoyet [contact], Benjamin Niay, Anne-Hélène Olivier.

Synthesizing walking motions that look realistic and diverse is a challenging task in animation, and even more when the target is to create realistic motions for large group of characters. Indeed, in order to keep a good trade-off between computational costs and realism, biomechanical constraints of human walk are not always fulfilled. In pilot experiments [38], [46], we have therefore started to investigate the ability of viewers to identify an invariant parameter of human walking named the walk ratio, representing the ratio between step length and step frequency of an individual, when applied to virtual humans. To this end, we recorded 4 actors (2 males, 2 females) walking at different freely chosen speeds, as well as at different combinations of step frequency and step length. We then performed pilot perceptual studies to identify the ability of viewers to detect the range of walk ratios considered as natural and compared it to the walk ratio freely chosen by the actor when performing walks at the same speeds. Our results will provide new considerations to drive the animation of walking virtual characters using the walk ratio as a parameter, which we believe could enable animators to control the speed of characters through simple parameters while retaining the naturalness of the locomotion.

7.3. Fidelity of Virtual Reality

MimeTIC wishes to promote the use of Virtual Reality to analyze and train human motor performance. It raises the fundamental question of the transfer of knowledge and skills acquired in VR to real life. In 2019, we put efforts in better understanding the potential fidelity of Virtual Reality experiences compared to real life experiences. It has been applied to various aspects of the interaction between pedestrians, but also the biomechanical fidelity of using haptic devices in highly constrainted conditions, such as hammering tasks.

7.3.1. Influence of Motion Speed on the Perception of Latency in Avatar Control

Participants: Ludovic Hoyet [contact], Richard Kulpa, Anthony Sorel, Franck Multon.

With the dissemination of Head Mounted Display devices in which users cannot see their body, simulating plausible avatars has become a key challenge. For fullbody interaction, avatar simulation and control involves several steps, such as capturing and processing the motion (or intentions) of the user using input interfaces, providing the resulting user state information to the simulation platform, computing a plausible adaptation of the virtual world, rendering the scene, and displaying the multisensory feedback to the user through output interfaces. All these steps imply that the displayed avatar motion appears to users with a delay (or latency) compared to their actual performance. Previous works have shown an impact of this delay on the perception-action loop, with possible impact on Presence and embodiment. We have explored [37] how the speed of the motion performed when controlling a fullbody avatar can impact the way people perceive and react to such a delay. We conducted an experiment where users were asked to follow a moving object with their finger, while embodied in a realistic avatar. We artificially increased the latency by introducing different levels of delays (up to 300ms) and measured their performance in the mentioned task, as well as their feeling about the perceived latency. Our results show that motion speed influenced the perception of latency: we found critical latencies of 80ms for medium and fast motion speeds, while the critical latency reached 120ms for a slow motion speed. We also noticed that performance is affected by both latency and motion speed, with higher speeds leading to decreased performance. Interestingly, we also found that performance was affected by latency before the critical latency for medium and fast speeds, but not for a slower speed. These findings could help to design immersive environments to minimize the effect of latency on the performance of the user, with potential impacts on Presence and embodiment.

7.3.2. Influence of Personality Traits and Body Awareness on the Sense of Embodiment in Virtual Reality

Participants: Ludovic Hoyet [contact], Rebecca Fribourg, Diane Dewez.

With the increasing use of avatars (i.e. the virtual representation of the user in a virtual environment) in virtual reality, it is important to identify the factors eliciting the sense of embodiment or the factors that can disrupt this feeling. This paper [35] reports an exploratory study aiming at identifying internal factors (personality traits and body awareness) that might cause either a resistance or a predisposition to feel a sense of embodiment towards a virtual avatar. To this purpose, we conducted an experiment (n=123) in which participants were immersed in a virtual environment and embodied in a gender-matched generic virtual avatar through a head-mounted display. After an exposure phase in which they had to perform a number of visuomotor tasks (during 2 minutes) a virtual character entered the virtual scene and stabbed the participants' virtual hand with a knife. The participants' sense of embodiment was measured, as well as several personality traits (Big Five traits and locus of control) and body awareness, to evaluate the influence of participants' personality on the acceptance of the virtual body. The major finding of the experiment is that the locus of control is linked to several components of embodiment: the sense of agency is positively correlated with an internal locus of control and the sense of body ownership is positively correlated with an external locus of control. Interestingly, both components are not influenced by the same traits, which confirms that they can appear independently. Taken together our results suggest that the locus of control could be a good predictor of the sense of embodiment when the user embodies an avatar with a similar physical appearance.

7.3.3. Gaze Behaviour During Person-Person Interaction in VR

Participants: Ludovic Hoyet, Anne-Hélène Olivier [contact], Florian Berton.

Figure 4. Conditions used in this work to understand the effect of VR setup on gaze beaviour in a collision avoidance task.

Simulating realistic interactions between virtual characters has been of interest to research communities for years, and is particularly important to automatically populate virtual environments. This problem requires to accurately understand and model how humans interact, which can be difficult to assess. In this context, Virtual Reality (VR) is a powerful tool to study human behaviour, especially as it allows assessing conditions which are both ecological and controlled. While VR was shown to allow realistic collision avoidance adaptations, in the frame of the ecological theory of perception and action, interactions between walkers can not solely be characterized through motion adaptations but also through the perception processes involved in such interactions. The objective of this study [30] is therefore to evaluate how different VR setups influence gaze behaviour during collision avoidance tasks between walkers. In collaboration with Julien Pettré in Rainbow team, we designed an experiment involving a collision avoidance task between a participant and another walker (real confederate or virtual character). During this task, we compared both the participant's locomotion and gaze behaviour in a real environment and the same situation in different VR setups (including a CAVE, a screen and a Head-Mounted Display) as illustrated on Figure4 . Our results show that even if some quantitative differences exist, gaze behaviour is qualitatively similar between VR and real conditions. Especially, gaze behaviour in VR setups including a HMD is more in line with the real situation than the other setups. Furthermore, the outcome on motion adaptations confirms previous work, where collision avoidance behaviour is qualitatively similar in VR and real conditions. In conclusion, our results show that VR has potential for qualitative analysis of locomotion and gaze behaviour during collision avoidance. This opens perspectives in the design of new experiments to better understand human behaviour, in order to design more realistic virtual humans.

7.3.4. Gaze Anticipation in Curved Path in VR

Participants: Anne-Hélène Olivier [contact], Hugo Brument.

This work was performed in collaboration with Ferran Argelaguet-Sanz and Maud Marchal from Hybrid team [32]. We investigated whether the body anticipation synergies in real environments (REs) are preserved during navigation in virtual environments (VEs). Experimental studies related to the control of human locomotion in REs during curved trajectories report a top-down reorientation strategy with the reorientation of the gaze anticipating the reorientation of head, the shoulders and finally the global body motion. This anticipation behavior provides a stable reference frame to the walker to control and reorient whole-body according to the future direction. To assess body anticipation during navigation in VEs, we conducted an experiment where participants, wearing a head-mounted display, were asked to perform a lemniscate trajectory in a virtual environment (VE) using five different navigation techniques, including walking, virtual steering (hand, head or torso steering) and passive navigation. For the purpose of this experiment, we designed a new control law based on the powerlaw relation between speed and curvature during human walking. Taken together our results showed that a similar ordered top-down sequence of reorientation of the gaze, head and shoulders during curved trajectories between walking in REs and in VEs (for all the evaluated techniques). However, this anticipation mechanism significantly differs between physical walking in VE, where the anticipation is higher, and the other virtual navigation techniques. The results presented in this paper pave the way to the better

understanding of the underlying mechanisms of human navigation in VEs and to the design of navigation techniques more adapted to humans

7.3.5. Validity of VR to Study Social Norms During Person-Person Interaction

Participants: Anne-Hélène Olivier [contact], Ludovic Hoyet, Florian Berton.

The modelling of virtual crowds for major events, such as the Oympics in Paris in 2024, takes into account the global proxemics standards of individuals without questioning the possible variability of these standards according to the space in which the interactions are performed. We know that body interactions (Goffman, 1974) are subject to rules whose variability is, at least in part, cultural (Hall, 1971). Obviously, these proxemics standards also address practical issues such as available space and space occupancy density. Our objective in this study was to understand the conditions which can explain that the discomfort felt and the adaptive behaviour performed differ when the interaction takes place in the same city and in spaces with identical occupancy densities. Especially, we focused on the effect of the social context of the environement. We aim at estimating the extent to which the prospect of attending a sports performance alters sensitivity to the transgression of proxemics norms. An additional objective was to evaluate whether virtual reality can help us to provide new insights in such a social context, where objective measures out-of-the lab are complex to perform. To answer this question, we designed in collaboration with Julien Pettré (Rainbow team) and colleagues in the field of sociology François Le Yondre, Théo Rougant and Tristan Duverne (Univ Rennes II) an experiment (in real context and then in virtual reality) in two different locations: a train station and the surroundings of a stadium before a league 1 football match) but with similar densities. The task performed by a confederate was to walk and stand excessively close to men aged 20 to 40. The individual's behaviour (not conscious of being a subject of the experiment) was observed by ethnography and explanatory interviews were conducted immediately afterwards. This same experiment was carried out in virtual reality conditions on the same type of population, modelling the two spaces and making it possible to acquire more precise and quantifiable data than in real conditions such as distances, travel time and eye fixations. The results show that the discomfort shown is much higher in the train station. The sporting context seems to participate in a form of relaxation of the norms of bodily interaction. Such a gap is not observable in virtual reality. From a methodological point of view, explicit interviews make it possible to usefully identify the reasons why virtual reality does not generate the same reactions, although it sometimes provokes the same sensitivity. Future work is needed to evaluate the effect of an increased immersion on such Social Science studies.

7.4. Motion Sensing of Human Activity

MimeTIC has a long experience in motion analysis in laboratory condition. In the MimeTIC project, we proposed to explore how these approaches could be transferred to ecological situations, with a lack of control on the experimental conditions. In the continuation of 2018, we have proposed to explore the use of cheap depth cameras solution for on-site motion analysis in ergonomics.

7.4.1. Motion Analysis of Work Conditions Using Commercial Depth Cameras in Real Industrial Conditions

Participant: Franck Multon [contact].

Based on a former PhD thesis (of Pierre Plantard) we have demonstrated the use of depth sensors in industry to assess risks of musculoskeletal disorders ar work. It has leaded to the creation of the the KIMEA software and of the Moovency start-up company in November 2018. In 2019 we published a synthesis work with new results [48] to demonstrate that such an approach can actually support the work of ergonomists in their goal to enhance the quality of life of workers in industry.

Hence, measuring human motion activity in real work condition is challenging as the environment is not controlled, while the worker should perform his/her task without perturbation. Since the early 2010s, affordable and easy-to-use depth cameras, such as the Microsoft Kinect system, have been applied for in-home entertainment for the general public. In this work, we evaluated such a system for the use in motion analysis in work conditions and propose software algorithms to enhance the tracking accuracy. Firstly, we highlighted the

high performance of the system when used under the recommended setup without occlusions. However, when the position/orientation of the sensor changes, occlusions may occur and the performance of the system may decrease, making it difficult to be used in real work conditions. Secondly, we propose a software algorithm to adapt the system to challenging conditions with occlusions to enhance the robustness and accuracy. Thirdly, we show that real work condition assessment using such an adapted system leads to similar results comparing with those performed manually by ergonomists. These results show that such adapted systems could be used to support the ergonomists work by providing them with reproducible and objective information about the human movement. It consequently saves ergonomists time and effort and allows them to focus on high-level analysis and actions.

7.5. Sports

MimeTIC promotes the idea of coupling motion analysis and synthesis in various domains, especially sports. More specifically, we have a long experience and international leadership in using Virtual Reality for analyzing and training sports performance. In 2019, we continued to explore 1) how enhancing on-site sports motion analysis using models inspired from motion simulation techniques, and 2) how Virtual Reality could be used to analyze and train motor and perceptual skills in sports.

7.5.1. Analysis of Fencing Lunge Accuracy and Response Time in Uncertain Conditions With an Innovative Simulator

Participants: Anthony Sorel [contact], Richard Kulpa, Nicolas Bideau, Charles Pontonnier.

We conducted a study evaluating the motor control strategies implied by the introduction of uncertainty in the realization of lunge motions [27]. Lunge motion is one of the fundamental attacks used in modern fencing, asking for a high level of coordination, speed and accuracy to be efficient. The aim of the current paper was the assessment of fencer's performance and response time in lunge attacks under uncertain conditions. For this study, an innovative fencing lunge simulator was designed. The performance of 11 regional to national-level fencers performing lunges in Fixed, Moving and Uncertain conditions was assessed. The results highlighted notably that i) Accuracy and success decreased significantly in Moving and Uncertain conditions with regard to Fixed ones ii) Movement and Reaction times were also affected by the experimental conditions iii) Different fencer profiles were distinguishable among subjects. In conclusion, the hypothesis that fencers may privilege an adaptation to the attack conditions and preserve accuracy instead of privileging quickness was supported by the results. Such simulators may be further used to analyze in more detail the motor control strategies of fencers through the measure and processing of biomechanical quantities and a wider range of fencing levels. It has also a great potential to be used as training device to improve fencer's performance to adapt his attack to controlled opponent's motion.

7.5.2. Enactive Approach to Assess Perceived Speed Error during Walking and Running in Virtual Reality

Participants: Théo Perrin, Richard Kulpa [contact], Charles Faure, Anthony Sorel, Benoit Bideau.

The recent development of virtual reality (VR) devices such as head mounted displays (HMDs) increases opportunities for applications at the confluence of physical activity and gaming. Recently, the fields of sport and fitness have turned to VR, including for locomotor activities, to enhance motor and energetic resources, as well as motivation and adherence. For example, VR can provide visual feedbacks during treadmill running, thereby reducing monotony and increasing the feeling of movement and engagement with the activity. However, the relevance of using VR tools during locomotion depends on the ability of these systems to provide natural immersive feelings, specifically a coherent perception of speed. The objective of this study is to estimate the error between actual and perceived locomotor speed in VE using an enactive approach, i.e. allowing an active control of the environment. Sixteen healthy individuals participated in the experiment, which consisted in walking and running on a motorized treadmill at speeds ranging from 3 to 11 km/h with 0.5 km/h increments, in a randomized order while wearing a HMD device (HTC Vive) displaying a virtual racetrack. Participants were instructed to match VE speed with what they perceived was their actual

locomotion speed (LS), using a handheld Vive controller. They were able to modify the optic flow speed (OFS) with a 0.02 km/h increment/decrement accuracy. An optic flow multiplier (OFM) was computed based on the error between OFS and LS. It represents the gain that exists between the visually perceived speed and the real locomotion speed experienced by participants for each trial. For all conditions, the average of OFM was 1.00 $\pm$.25 to best match LS. This finding is at odds with previous works reporting an underestimation of speed perception in VR. It could be explained by the use of an enactive approach allowing an active and accurate matching of visually and proprioceptively perceived speeds by participants. But above all, our study showed that the perception of speed in VR is strongly individual, with some participants always overestimating and others constantly underestimating. Therefore, a general OFM should not be used to correct speed in VE to ensure congruence in speed perception, and we propose the use of individual models as recommendations for setting up locomotion-based VR applications.

7.5.3. *Acting Together, Acting Stronger? Interference Between Participants During Face-to-Face Cooperative Interception Task*

Participants: Charles Faure, Théo Perrin, Richard Kulpa [contact], Anthony Sorel, Anabelle Limballe, Benoit Bideau.

People generally coordinate their action to be more effective. However, in some cases, interference between them occur, resulting in an inefficient collaboration. The main goal of this study [16], [16] is to explore the way two persons regulate their actions when performing a cooperative task of ball interception, and how interference between them may occur. Starting face to face, twenty-four participants (twelve teams of two) had to physically intercept balls moving down from the roof to the floor in a virtual room. To this end, they controlled a virtual paddle attached to their hand moving along the anterior-posterior axis, and were not allowed to communicate. Results globally showed participants were often able to intercept balls without collision by dividing the interception space in two equivalent parts. However, an area of uncertainty (where many trials were not intercepted) appeared in the center of the scene highlighting the presence of interference between participants. The width of this area increased when situation became more complex and when less information was available. Moreover, participants often interpreted balls starting above them as balls they should intercept, even when these balls were in fine intercepted by their partner. Overall, results showed that team coordination emerges from between-participants interactions in this ball interception task and that interference between them depends on task complexity (uncertainty on partner's action and visual information available).

7.5.4. *Detection of Deceptive Motions in Rugby from Visual Motion Cues*

Participants: Richard Kulpa [contact], Anne-Hélène Olivier, Benoit Bideau.

Figure 5. Illustration of a rugby player interacting in VR with 3 representations of a virtual attacker.

Frequently, in rugby, players incorporate deceptive motions (e.g., a side-step) in order to pass their opponent. Previous works showed that expert defenders are more efficient in detecting deceptive motions. Performance was shown to be correlated with the evolution of the center of gravity of the attacker, suggesting that experts may rely on global motion cues. This study [19] aims at investigating whether a representation of center of gravity can be useful for training purposes, by using this representation alone or by combining it with the local motion cues given by body parts. We designed an experiment in virtual reality to control the motion cues available to the defenders. Sixteen healthy participants (seven experts and nine novices) acted as defenders while a virtual attacker approached. Participants completed two separate tasks. The first was a time occlusion perception task, occlusion after 100ms, 200ms or 300ms after the initial change in direction, thereafter participants indicated the passing direction of the attacker. The second was a perception-action task Figure 5 , participants were instructed to intercept the oncoming attacker by displacing medio-laterally. The attacker performed either a non-deceptive motion, directly toward the final passing direction or a deceptive motion, initially toward a false direction before quickly reorienting to the true direction. There was a main effect of expertise, appearance, cut off times and motion on correct responses during both tasks. There was an interaction between visual appearance and expertise, and between motion type and expertise during the perception task, however, this interaction was not present during the perception-action task. We observed that experts maintained superiority in the perception of deceptive motion; however when the visual appearance is reduced to global motion alone the difference between novices and experts is reduced. We further explore the interactions and discuss the effects observed for the visual appearance and expertise.

7.5.5. IMU-based Motion Capture for Cycling Performance

Participants: Nicolas Bideau [contact], Guillaume Nicolas, Benoit Bideau, Sebastien Cordillet, Erwan Delhaye.

The quantification of 3D kinematical parameters such as body segment orientations and joint angles is important in the monitoring of cycling to provide relevant biomechanical parameters associated with performance optimization and/or injury prevention. Numerous experiments based on optoelectronic motion capture have been conducted in the laboratory to analyze kinematical variables (e.g., joint angles) during cycling. However, the assessment of kinematics in real conditions during training or competition is a challenging task, especially since conventional optoelectronic motion capture systems suffer from major drawbacks (restricted fields of view, cumbersome and time consuming) in this regard. To overcome these limitations, inertial measurement units (IMU) is a relevant solution for in situ cycling analysis as they allow a continuous data acquisition process throughout a cycling exercise. Beyond the common problem of the drift related to the integration of gyroscope data, one of the major issues in joint kinematics assessment using IMU devices lies is the misalignment of sensor axes with the anatomical body segment axis, which is not straightforward. Thus, we developed a novel sensor-to-segment calibration procedure for inertial sensor-based knee joint kinematics analysis during cycling. This procedure was designed to be feasible in-field, autonomously, and without any external operator or device. It combines a static standing up posture and a pedaling task. In comparison with conventional calibration methods commonly employed in gait analysis, the new method we proposed significantly improved the accuracy of 3D knee joint angle measurement when applied to cycling analysis [14]. As a second step related to the in-field application to track cycling, we estimated lower limb joint angles during a time trial on a velodrome. This integrative measurement exhibited the evolution of kinematic parameters in relation with distance but also with the track curvature [43].

7.6. Ergonomics

Ergonomics has become an important application domains in MimeTIC: being able to capture, analyze, and model human performance at work. In this domain, key challenge consists in using limited equipment to capture the physical activity of workers in real conditions. Hence, in 2019, we have designed a new approach to predict external forces using mainly motion capture data, and to personnalize the biomechanical capabilities (maximum feasible force/torque) of specific population.

7.6.1. Motion-based Prediction of External Forces

Participants: Charles Pontonnier [contact], Georges Dumont, Claire Livet, Anthony Sorel, Nicolas Bideau.

We proposed [21] a method to predict the external efforts exerted on a subject during handling tasks, only with a measure of his motion. These efforts are the contacts forces and moments on the ground and on the load carried by the subject. The method is based on a contact model initially developed to predict the ground reaction forces and moments. Discrete contact points are defined on the biomechanical model at the feet and the hands. An optimization technique computes the minimal forces at each of these points satisfying the dynamic equations of the biomechanical model and the load. The method was tested on a set of asymmetric handling tasks performed by 13 subjects and validated using force platforms and an instrumented load. For each task, predictions of the vertical forces obtained a RMSE of about 0.25 N/kg for the feet contacts and below 1 N/kg for the hands contacts. This method enables to quantitatively assess asymmetric handling tasks on the basis of kinetics variables without additional instrumentation such as force sensors and thus improve the ecological aspect of the studied tasks. We evaluated this method [23] on manual material handling (MMH) tasks. From a set of hypothesized contact points between the subject and the environment (ground and load), external forces were calculated as the minimal forces at each contact point while ensuring the dynamics equilibrium. Ground reaction forces and moments (GRF&M) and load contact forces and moments (LCF&M) were computed from motion data alone. With an inverse dynamics method, the predicted data were then used to compute kinetic variables such as back loading. On a cohort of 65 subjects performing MMH tasks, the mean correlation coefficients between predicted and experimentally measured GRF for the vertical, antero-posterior and medio-lateral components were 0.91 (0.08), 0.95 (0.03) and 0.94 (0.08), respectively. The associated RMSE were 0.51 N/kg, 0.22 N/kg and 0.19 N/kg. The correlation coefficient between L5/S1 joint moments computed from predicted and measured data was 0.95 with a RMSE of 14 Nm for the flexion / extension component. This method thus allows the assessment of MMH tasks without force platforms, which increases the ecological aspect of the tasks studied and enables performance of dynamic analyses in real settings outside the laboratory.

This method was successfully applied [24] on lunge motion that is a fundamental attack of modern fencing, asking for a high level of coordination, speed and accuracy. It consists in an explosive extension of the front leg accompanying an extension of the sword arm. In such motions, the direction of action and the way feet are oriented – guard position - are particularly challenging for a GRF&M prediction method. These methods are avalaible in CusToM software [22].

7.6.2. Biomechanics for Motion Analysis-Synthesis and Analysis of Torque Generation Capacities

Participants: Charles Pontonnier [contact], Georges Dumont, Nicolas Bideau, Guillaume Nicolas, Pierre Puchaud.

Characterization of muscle mechanism through the torque-angle and torque-velocity relationships [17] is critical for human movement evaluation and simulation. In-vivo determination of these relationships through dynamometric measurements and modelling is based on physiological and mathematical aspects. However, no investigation regarding the effects of the mathematical model and the physiological parameters underneath these models was found. The purpose of the current study was to compare the capacity of various torque-angle and torque-velocity models to fit experimental dynamometric measurement of the elbow and provide meaningful mechanical and physiological information. Therefore, varying mathematical function and physiological muscle parameters from the literature were tested. While a quadratic torque-angle model seemed to increase predicted to measured elbow torque fitting, a new power-based torque-velocity parametric model gave meaningful physiological values with similar fitting results to a classical torque-velocity model. This model is of interest to extract modelling and clinical knowledge characterizing the mechanical behavior the joint. Based on the same kind of methods, we proposed [25] to analyse torque generation capacities of a human knee. The torque generation capacities are often assessed for human performance, as well as for prediction of internal forces through musculoskeletal modelling. Scaling individual strength generation capacities is challenging but can provide physiologically meaningful perspectives. We propose to fit the models to isokinetic measurements of joint torques in different angle and angular velocity conditions. Assuming muscles are viscoelastic actuators, their entire architectures contribute to Joint Torque-Angle and Torque-Velocity Relationships (JTAR and JTVR respectively, and their coupling JTAVR) at the joint level. Experimental observation at different scales (muscle sarcomere, muscle fibre and joint) resulted in various JTAR models available in the literature. On

the other side, JVTR models are often modelled without obvious physiological consistency. The above mentionned JTVR model was shown to increase physiological transparency of the elbow JTAVR. As those results might be joint-specific, we extended it to evaluate five JTAR and two JTVR models on the knee flexion and extension.

7.7. Locomotion and Interactions between Walkers

MimeTIC is a leader in the study and modeling of walkers' visuo-motor strategies. This implies to understand how humans generate their walking trajectories within an environment. This year, one main focus was to consider how the interaction models change with specific populations (including kids, older adults, concussed athletes or person on a wheelchair) as well as in specific environment (including narrow sidewalk, or environment with varying social context).

7.7.1. Effect of Foot Stimulation on Locomotion

Participants: Anne-Hélène Olivier, Armel Crétual [contact], Carole Puil.

Medio-Intern Element (EMI®) is a thin plantar insert used by podiatrists to treat postural deficiency. It was shown an influence of a 3 mm high EMI on Medio-Lateral (ML) displacement of the Centre of Pressure (CoP) of healthy participants in quasi-static standing. Recently it has been demonstrated that EMI has an impact on eyes vergence, and especially in population with plantar postural dysfunction. These effects were weakly assessed however and only using static tasks. Therefore, the objective of this work [53], [52], [41], was to evaluate the effect of the EMI while performing a locomotor task. We expected a controlateral deviation of the trajectory when this insert was located under one foot. Indeed, in previous studies dealing with bottom-up control of locomotion, it was shown that a 30 min podokinetic stimulation leads to a ML deviation of the trajectory when participants were asked to walk in a straight line with eyes closed. 20 healthy participants volunteered for this study. They participated into 3 different sessions in random order: either without EMI, with EMI under the right foot or under the left foot. Each session involved first, static tasks (with and without vision) to compare with previous work, then, dynamic locomotor tasks with 6 different conditions mixing trajectory (straight walking, 90° left or right turn) and vision (with and without vision) in random order. In static conditions, we computed the average ML position of the CoP. In dynamic conditions, we analyzed the difference in the final orientation of the locomotor trajectory with and without vision with an EMI with respect to this difference without the EMI. No significant effect of the EMI was observed for either static or dynamic conditions. Our results do not confirm the previous work in static conditions. Future work is needed to better understand the effect of this insert. In particular, our participants were healthy and it could be interesting to evaluate this effect in participants with postural deficiencies. These results would have an application in the design of new clinical tests.

7.7.2. Collision Avoidance between Walkers on a Curvilinear Path

Participants: Anne-Hélène Olivier [contact], Armel Crétual, Richard Kulpa, Anthony Sorel.

Crowded public spaces require humans to interact with what the environment affords to regulate interpersonal distance to avoid collisions. In the case of rectilinear trajectories, the collision avoidance behaviours have been extensively studied. It has been shown that the perceived action-opportunities of the walkers might be afforded based on a future distance of closest approach (also coined 'Minimal Predicted Distance', MPD). However, typical daily interactions do not always follow rectilinear but also curvilinear trajectories. In that context, it has been shown that a ball following a curvilinear trajectory can be successfully intercepted. However, it remains unclear whether the collision avoidance strategies in the well-studied linear trajectories can be transferred to curvilinear trajectories. Therefore, the aim of this work [44] was to examine collision avoidance behaviours when interacting with walkers following curvilinear trajectories. An experiment was designed using virtual reality in which 22 participants navigated toward a goal in a virtual environment with a joystick. A Virtual Human (VH) crossed the path of the participant from left and right with varying risks of collision. The VH followed either a curvilinear path with a fixed radius of 5 m or 10 m, approaching from in-front of and behind the participant, or a control rectilinear path. The final crossing distance, the number of collisions and inversions

of initial crossing order were analysed to determine the success of the task. Further, MPD evolution over time and specific timing events was analysed across conditions. For a curvilinear path with a 5 m radius there were significantly more collisions when the VH approached from behind the participant, and significantly more inversions of the initial crossing order when the VH approached from in-front than the control rectilinear path. Final crossing distance was shorter when the VH followed a path with a 5 m radius from behind the participant. Finally, the evolution of the MPD over time was similar for paths with a 10 m radius when compared to the control rectilinear path, whereas the 5 m curvilinear paths had significant differences during the interaction. Overall, with few collisions and few inversions of crossing order we can conclude that participants were capable of interacting with virtual walkers on curvilinear trajectories. Further, the task was solved with similar avoidance adaptations to those observed for rectilinear interactions. However, paths with a smaller radius had more reported collisions and inversions. Future work should address how a curved trajectory during collision avoidance is perceived.

7.7.3. Collision Avoidance in Person-Specific Populations

Participants: Anne-Hélène Olivier [contact], Armel Crétual.

In the frame of the Inria BEAR associate team, we have used our 90° crossing paradigm to understand visuo-motor coordination in specific population. This is important, not only from a theoretical point of view but also to design more individual model of human locomotion in a dynamic environment.

Figure 6. Illustration of person-person interaction experiments in a) kids, b) older adults, c) previously concussed athletes, d) a person on an electric powered wheelchair

We first investigated the effect of age on visuo-motor coordination by considering a collision avoidance task in kids (8-12 years) and older adults (65-74 years) as illustrated on Figure 6 a,b. On one hand, middle-aged children have been shown to have poor perception-action coupling during static and dynamic collision avoidance tasks. Research has yet to examine whether perception-action coupling deficits persist in a dynamic collision avoidance task involving a child and another walker. In this work [26], [54], we invesigated whether the metric MPD(t) be used to examine collision avoidance strategies between children and adults. To this end, eighteen children (age: 10 $\pm$ 1.5 years) and eighteen adults (34 $\pm$ 9.6 years) walked while avoiding another participant (child or adult). Groups of three children and three adults were recruited per session. The results demonstrated that (1) MPD(t) can be used to predict future collisions in children, (2) MPD(t) is an absolute measure that is consistently lower when a child is involved compared to two adult walkers, (3) the individual passing second, even when it is a child, contributes more to MPD(t) than the walker passing first. It then appears that children have developed adult-like strategies during a collision avoidance task involving two walkers. Body anthropometrics should be considered when determining collision avoidance strategies between children and adults. On the other hand, every year, 1 in 3 older adults are likely to fall at least once and many falls occurs while walking where an individual needs to adapt to environmental hazards. Studies with older adults interacting within an environment showed difficulties in estimating time to arrival of vehicles, larger critical ratio and more variability in door aperture task as well as larger clearance distance when avoiding a moving object. The current study [51] aims to identify whether differences in collision

avoidance behaviours of older adults during a person-person collision avoidance task are the result of age-related visuomotor processing deficits. Results showed that no collision occurred, where older and younger adults were able to act appropriately. However, larger threshold were needed to trigger avoidance when an older adult is second in crossing order, possibly due to visuomotor delays. Moreover, we observed more crossing inversions with older adults, which may suggest a poor visumotor processing. Finally, the clearance distance was samller when to older adults interact with each others, resulting in “risky” behaviours. Interestingly, social factors seems to be involved since when a young and an older adults interact, the young adult contribute more to solve the collision avoidance task.

In close relation with the Application Domain "Sports", we also investigated visuo-motor coordination during locomotion in previously concussed rugby-players (Figure 6 c). Despite adherence to return-to-play guidelines, athletes with previous concussion exhibit persistent visuomotor deficits during static balance and visuomotor integration tasks such as collision avoidance months after returning to sport. Previous research in collision avoidance was done in a static setting, however less is known about visuomotor strategies utilized in dynamic scenarios, such as person-person interactions. In this context, during a collision avoidance locomotor task, individuals make adjustments to their path and/or their velocity in response to a risk of collision. These adjustments ensure that the clearance distance would be large enough such that no collision occurs. However, athletes with previous concussion may demonstrate impaired performance during a collision avoidance task requiring path adjustments based on visual information. The purpose of this study [55] was to investigate collision avoidance strategies when avoiding another walker between previously concussed athletes and healthy athletes. We hypothesized that previously concussed athletes would demonstrate altered trajectory adaptation and changes in individual contribution to the avoidance compared to healthy athletes. Preliminary results show that individuals with previous concussion demonstrated trajectory adaptation behaviours consistent with healthy athletes and young adults. However, previously concussed athletes passed with a reduced distance between themselves and the other walker when they are second in passage order at the crossing point. Athletes who have sustained a previous concussion show decreased collision avoidance behaviour. This behaviour results in a higher risk of a collision occurring, as individuals showed reduced contributions (i.e. creating physical space) to the avoidance of the collision. This change in typical behaviour on a visuomotor task may indicate a persistent deficit in perceptual abilities following concussion. Although trajectory adaptations were consistent with healthy athletes, these results suggest that athletes with previous concussion remain at an elevated risk of collision and possible injury following concussion recovery. This study provides novel insights and additional evidence that visuomotor and perceptual impairments persist following return to play in previously concussed athletes. Additionally, this protocol has important implications for the assessment and rehabilitation of visuomotor processes that are affected following a concussion. Future research could further develop this protocol to be used in sideline assessment, and guide treatment of concussions past clinical recovery.

7.7.4. Collision Avoidance between a Walker and an Electric Powered Wheelchair: Towards Smart Wheelchair

Participants: Anne-Hélène Olivier [contact], Armel Crétual.

In collaboration with Marie Babel and Julien Pettré from Inria Rainbow team, we are interested in the development of smart electric powered wheelchairs (EPW), which provide driver assistance. Developing smart assistance requires to better understand interactions between walkers and such vehicles. We focus on collision avoidance task between an EPW (fully operated by a human) and a walker, where the difference in the nature of the agents (weight, maximal speed, acceleration profiles) results into asymmetrical physical risk in case of a collision, for example due to the protection EPW provides to its driver, or the higher energy transferred to the walker during head-on collision. In this work [39], [47], our goal is to demonstrate that this physical risk asymmetry results into differences in the walker’s behavior during collision avoidance in comparison to human-human situations. 20 participants (15 walkers and 5 EPW drivers) volunteered to this study. The experiment was performed in a 30mx20m gymnasium. We designed a collision avoidance task, where an EPW and a human walker moved towards a goal with orthogonal crossing trajectories (Figure6 d). We recorded their trajectory among 246 trials (each trial being 1 collision avoidance). We compared the predicted passage order

when they can first see each other with the one observed at the crossing point to identify if inversions occur during the interaction. Note that during walker-walker interactions it was shown that the initial passage order is almost systematically preserved all along the interaction up to the crossing point. We also computed the shape-to-shape clearance distance. We observed 23.7% of passage order inversion, specifically in 20.8% of trials where walkers were supposed to cross first, they crossed second. This means that walkers were more likely to pass behind the EPW than in front. On average, human walkers crossed first when having sufficient advance on the wheelchair to reach the crossing point. We estimated this advance up to 0.91m. The shape-to-shape clearance distance was influenced by the passage order at the crossing point, with larger distance when the walker cross first (M=0.78m) than second (M=0.34m). Results show that walkers set more conservative strategies when interacting with an EPW. By passing more frequently behind the EPW, they avoid risks of collisions that would lead to high energy transfer. Also, when they pass in front, they significantly increase the clearance distance, compared to cases where they pass behind. These results can then be linked to the difference in the physical characteristics of the walkers and EPW where asymmetry in the physical risks raised by collisions influence the strategies performed by the walkers in comparison with a similar walker-walker situation. This gives interesting insights in the task of modeling such interactions, indicating that geometrical terms are not sufficient to explain behaviours, physical terms linked to collision momentum should also be considered.

7.7.5. Collision Avoidance on a Narrow Sidewalk

Participant: Anne-Hélène Olivier [contact].

In the context of transportation research and a collaboration with the colleagues of Ifsttar (LEPSIS, LESCOT), we investigate person-person interaction when walking on a narrow sidewalk [34]. Narrow sidewalks are not the result of imagination nor a heritage of the former urban planning in the oldest cities. They exist in many modern cities, a simple web query provides a lot of examples in the world. In most cases, two pedestrians walking in opposite way cannot stay both on the sidewalk when they cross: one has to give a free way on the curb by stepping down on the road, which can generate risky situations for pedestrians. These situations are nowadays underestimated and so are the associated risk. In this context, driving simulators and walking simulators are useful tools to conduct studies in a safe environment with controlled conditions. Therefore, they can allow improving our knowledge on the way pedestrians interact on a narrow sidewalk and how drivers can react when facing this situation. This contribution aims to model the behaviours of simulated pedestrians, Non Player Characters (NPC). Using an interdisciplinary framework, we first identified from the literature psychosocial factors that should be involved in such interactions. Then, we designed a questionnaire to evaluate the impact of these factors on the perception of these interaction. Based on the main factors, we developed a perception model, and we modified the ORCA model, which is one of the most used for pedestrian collision avoidance simulation. Finally, we assessed the consistency of all our simulated interactions with a user study.

7.7.6. Shared Effort Model During Collision Avoidance

Participants: Anne-Hélène Olivier [contact], Armel Crétual.

In collaboration with Jose Grimaldo da Silva and Thierry Fraichard (Inria Grenoble), we finally designed a shared-effort model during interaction between a moving robot and a human relying on walker-walker collision avoidance data. [33]. Recent works in the domain of Human-Robot Motion (HRM) attempted to plan collision avoidance behavior that accounts for cooperation between agents. Cooperative collision avoidance between humans and robots should be conducted under several factors such as speed, heading and also human attention and intention. Based on some of these factors, people decide their crossing order during collision avoidance. However, whenever situations arise in which the choice crossing order is not consistent for people, the robot is forced to account for the possibility that both agents will assume the same role, a decision detrimental to collision avoidance. In our work we evaluate the boundary that separates the decision to avoid collision as first or last crosser. Approximating the uncertainty around this boundary allows our collision avoidance strategy to address this problem based on the insight that the robot should plan its collision avoidance motion in such a way that, even if agents, at first, incorrectly choose the same crossing order, they would be able to unambiguously perceive their crossing order on their following collision avoidance action.

PANAMA Project-Team

7. New Results

7.1. Sparse Representations, Inverse Problems, and Dimension Reduction

Sparsity, low-rank, dimension-reduction, inverse problem, sparse recovery, scalability, compressive sensing

The team's activity ranges from theoretical results to algorithmic design and software contributions in the fields of sparse representations, inverse problems and dimension reduction.

7.1.1. Computational Representation Learning: Algorithms and Theory

Participants: Rémi Gribonval, Hakim Hadj Djilani, Cássio Fraga Dantas, Jeremy Cohen.

Main collaborations: Luc Le Magoarou (IRT b-com, Rennes), Nicolas Tremblay (GIPSA-Lab, Grenoble), R. R. Lopes and M. N. Da Costa (DSPCom, Univ. Campinas, Brazil)

An important practical problem in sparse modeling is to choose the adequate dictionary to model a class of signals or images of interest. While diverse heuristic techniques have been proposed in the literature to learn a dictionary from a collection of training samples, classical dictionary learning is limited to small-scale problems. In our work introduced below, by imposing structural constraints on the dictionary and pruning provably unused atoms, we could alleviate the curse of dimensionality.

Multilayer sparse matrix products for faster computations. Inspired by usual fast transforms, we proposed a general dictionary structure (called FAμST for Flexible Approximate Multilayer Sparse Transforms) that allows cheaper manipulation, and an algorithm to learn such dictionaries together with their fast implementation, with reduced sample complexity. A comprehensive journal paper was published in 2016 [75], and we further explored the application of this technique to obtain fast approximations of Graph Fourier Transforms [76], empirically showing that $\mathcal{O}(n \log n)$ approximate implementations of Graph Fourier Transforms are possible for certain families of graphs. This opened the way to substantial accelerations for Fourier Transforms on large graphs. This year we focused on the development of the FAμST software library (see Section 6), providing transparent interfaces of FAμST data-structures with both Matlab and Python.

Kronecker product structure for faster computations. In parallel to the development of FAuST, we proposed another approach to structured dictionary learning that also aims at speeding up both sparse coding and dictionary learning. We used the fact that for tensor data, a natural set of linear operators are those that operate on each dimension separately, which correspond to rank-one multilinear operators. These rank-one operators may be cast as the Kronecker product of several small matrices. Such operators require less memory and are computationally attractive, in particular for performing efficient matrix-matrix and matrix-vector operations. In our proposed approach, dictionaries are constrained to belong to the set of low-rank multilinear operators, that consist of the sum of a few rank-one operators. The general approach, coined HOSUKRO for High Order Sum of Kronecker products, was shown last year to reduce empirically the sample complexity of dictionary learning, as well as theoretical complexity of both the learning and the sparse coding operations [67]. This year we demonstrated its potential for hyperspectral image denoising. A new efficient algorithm with lighter sample complexity requirements and computational burden was proposed and shown to be competitive with the state-of-the-art for hyperspectral image denoising with dedicated adjustements [50], [28], [27].

Combining faster matrix-vector products with screening techniques. We combined accelerated matrix-vector multiplications offered by FAμST / HOSUKRO matrix approximations with dynamic screening [59], that safely eliminates inactive variables to speedup iterative convex sparse recovery algorithms. First, we showed how to obtain safe screening rules for the exact problem while manipulating an approximate dictionary [68]. We then adapted an existing screening rule to this new framework and define a general procedure to leverage the advantages of both strategies. This lead to a journal publication [21] that includes new techniques based on duality gaps to optimally switch from a coarse dictionary approximation to a finer one. Significant complexity reductions were obtained in comparison to screening rules alone.

7.1.2. Generalized matrix inverses and the sparse pseudo-inverse

Participant: Rémi Gribonval.

Main collaboration: Ivan Dokmanic (University of Illinois at Urbana Champaign, USA)

We studied linear generalized inverses that minimize matrix norms. Such generalized inverses are famously represented by the Moore-Penrose pseudoinverse (MPP) which happens to minimize the Frobenius norm. Freeing up the degrees of freedom associated with Frobenius optimality enables us to promote other interesting properties. In a first part of this work [64], we looked at the basic properties of norm-minimizing generalized inverses, especially in terms of uniqueness and relation to the MPP. We first showed that the MPP minimizes many norms beyond those unitarily invariant, thus further bolstering its role as a robust choice in many situations. We then concentrated on some norms which are generally not minimized by the MPP, but whose minimization is relevant for linear inverse problems and sparse representations. In particular, we looked at mixed norms and the induced $\ell^p \to \ell^q$ norms.

An interesting representative is the sparse pseudoinverse which we studied in much more detail in a second part of this work published this year [19], motivated by the idea to replace the Moore-Penrose pseudoinverse by a sparser generalized inverse which is in some sense well-behaved. Sparsity implies that it is faster to apply the resulting matrix; well-behavedness would imply that we do not lose much in stability with respect to the least-squares performance of the MPP. We first addressed questions of uniqueness and non-zero count of (putative) sparse pseudoinverses. We showed that a sparse pseudoinverse is generically unique, and that it indeed reaches optimal sparsity for almost all matrices. We then turned to proving a stability result: finite-size concentration bounds for the Frobenius norm of p-minimal inverses for $1 \leq p \leq 2$. Our proof is based on tools from convex analysis and random matrix theory, in particular the recently developed convex Gaussian min-max theorem. Along the way we proved several results about sparse representations and convex programming that were known folklore, but of which we could find no proof.

7.1.3. Algorithmic exploration of large-scale Compressive Learning via Sketching

Participants: Rémi Gribonval, Antoine Chatalic.

Main collaborations this year: Nicolas Keriven (ENS Paris), Phil Schniter & Evan Byrne (Ohio State University, USA), Laurent Jacques & Vincent Schellekens (Univ Louvain, Belgium), Florimond Houssiau & Y.-A. de Montjoye (Imperial College London, UK)

Sketching for Large-Scale Learning. When learning from voluminous data, memory and computational time can become prohibitive. We proposed during the Ph.D. thesis of Anthony Bourrier [60] and Nicolas Keriven [74] an approach based on sketching. A low-dimensional sketch is computed by averaging (random) features over the training collection. The sketch can be seen as made of a collection of empirical generalized moments of the underlying probability distribution. Leveraging analogies with compressive sensing, we experimentally showed that it is possible to precisely estimate the mixture parameters provided that the sketch is large enough, and released an associated toolbox for reproducible research (see SketchMLBox, Section 6) with the so-called Compressive Learning Orthogonal Matching Pursuit (CL-OMP) algorithm which is inspired by Matching Pursuit. Three unsupervised learning settings have been addressed so far: Gaussian Mixture Modeling, k-means clustering, and principal component analysis. A survey conference paper on sketching for large-scale learning was published this year [25], and an extended journal version of this survey is in preparation.

Efficient algorithms to learn for sketches Last year, we showed that in the high-dimensional setting one can substantially speedup both the sketching stage and the learning stage with CL-OMP by replacing Gaussian random matrices with fast random linear transforms in the sketching procedure [63]. We studied an alternative to CL-OMP for cluster recovery from a sketch, which is based on simplified hybrid generalized approximate message passing (SHyGAMP). Numerical experiments suggest that this approach is more efficient than CL-OMP (in both computational and sample complexity) and more efficient than k-means++ in certain regimes [61]. During his first year of Ph.D., Antoine Chatalic visited the group of Phil Schniter to further investigate this topic, and a journal paper has been published as a result of this collaboration [15].

Privacy-preserving sketches Sketching provides a potentially privacy-preserving data analysis tool, since the sketch does not explicitly disclose information about individual datum. We established theoretical privacy guarantees (with the *differential privacy* framework) and explored the utility / privacy tradeoffs of Compressive K-means [24]. A journal paper is in preparation where we extend these results to Gaussian mixture modeling and principal component analysis.

Advances in optical-based random projections Random projections are a key ingredient of sketching. Motivated by the recent development of dedicated optics-based hardware for rapid random projections, which leverages the propagation of light in random media, we tackled the problem of recovering the phase of complex linear measurements when only magnitude information is available and we control the input. A signal of interest $\xi \in \mathbb{R}^N$ is mixed by a random scattering medium to compute the projection $y = \mathbf{A}\xi$, with $\mathbf{A} \in \mathbb{C}^{M\times N}$ a realization of a standard complex Gaussian independent and identically distributed (iid) random matrix. Such optics-based matrix multiplications can be much faster and energy-efficient than their CPU or GPU counterparts, yet two difficulties must be resolved: only the intensity $|y|^2$ can be recorded by the camera, and the transmission matrix $\mathbf{A}$ is unknown. We showed that even without knowing $\mathbf{A}$, we can recover the unknown phase of y for some equivalent transmission matrix with the same distribution as $\mathbf{A}$. Our method is based on two observations: first, conjugating or changing the phase of any row of $\mathbf{A}$ does not change its distribution; and second, since we control the input we can interfere ξ with arbitrary reference signals. We showed how to leverage these observations to cast the measurement phase retrieval problem as a Euclidean distance geometry problem. We demonstrated appealing properties of the proposed algorithm in both numerical simulations and real hardware experiments. Not only does our algorithm accurately recover the missing phase, but it mitigates the effects of quantization and the sensitivity threshold, thus improving the measured magnitudes [33].

7.1.4. Theoretical results on Low-dimensional Representations, Inverse problems, and Dimension Reduction

Participants: Rémi Gribonval, Clément Elvira, Jérémy Cohen.

Main collaboration: Nicolas Keriven (ENS Paris), Gilles Blanchard (Univ Postdam, Germany), Cédric Herzet (SIMSMART project-team, IRMAR / Inria Rennes), Charles Soussen (Centrale Supelec, Gif-sur-Yvette), Mila Nikolova (CMLA, Cachan), Nicolas Gillis (UMONS)

Information preservation guarantees with low-dimensional sketches. We established a theoretical framework for sketched learning, encompassing statistical learning guarantees as well as dimension reduction guarantees. The framework provides theoretical grounds supporting the experimental success of our algorithmic approaches to compressive K-means, compressive Gaussian Mixture Modeling, as well as compressive Principal Component Analysis (PCA). A comprehensive preprint is being revised for a journal [71].

Recovery guarantees for algorithms with continuous dictionaries. We established theoretical guarantees on sparse recovery guarantees for a greedy algorithm, orthogonal matching pursuit (OMP), in the context of continuous dictionaries [66], e.g. as appearing in the context of sparse spike deconvolution. Analyses based on discretized dictionary fail to be conclusive when the discretization step tends to zero, as the coherence goes to one. Instead, our analysis is directly conducted in the continuous setting amd exploits specific properties of the positive definite kernel between atom parameters defined by the inner product between the corresponding atoms. For the Laplacian kernel in dimension one, we showed in the noise-free setting that OMP exactly recovers the atom parameters as well as their amplitudes, regardless of the number of distinct atoms [66]. A preprint describing a full class of kernels for which such an analysis holds, in particular for higher dimensional parameters, has been released and submitted to a journal [30], [36], [31], [51].

Identifiability of Complete Dictionary Learning In the era of deep learning, dictionary learning has proven to remain an important and extensively-used data mining and processing tool. Having been studied and used for over twenty years, dictionary learning has well-understood properties. However there was a particular stone missing, which was understanding deterministic conditions for the parameters of dictionary learning to be uniquely retrieved from a training data set. We filled this gap partially by drastically improving on the previously best such conditions in the case of complete dictionaries [16]. Moreover, although algorithms with guaranties to compute the unique best solution do exist, they are seldom used in practice due to their

high computational cost. In subsequent work, we showed that faster algorithms typically used to compute dictionary learning often failed at computing the unique solution (in cases where our previous result guaranties this uniqueness), opening the way to new algorithms that are both fast and guarantied [26].

On Bayesian estimation and proximity operators. There are two major routes to address the ubiquitous family of inverse problems appearing in signal and image processing, such as denoising or deblurring. The first route is Bayesian modeling: prior probabilities are used to model both the distribution of the unknown variables and their statistical dependence with the observed data, and estimation is expressed as the minimization of an expected loss (e.g. minimum mean squared error, or MMSE). The other route is the variational approach, popularized with sparse regularization and compressive sensing. It consists in designing (often convex) optimization problems involving the sum of a data fidelity term and a penalty term promoting certain types of unknowns (e.g., sparsity, promoted through an L1 norm).

Well known relations between these two approaches have lead to some widely spread misconceptions. In particular, while the so-called Maximum A Posterori (MAP) estimate with a Gaussian noise model does lead to an optimization problem with a quadratic data-fidelity term, we disprove through explicit examples the common belief that the converse would be true. In previous work we showed that for denoising in the presence of additive Gaussian noise, for any prior probability on the unknowns, the MMSE is the solution of a penalized least-squares problem, with all the apparent characteristics of a MAP estimation problem with Gaussian noise and a (generally) different prior on the unknowns [72]. In other words, the variational approach is rich enough to build any MMSE estimator associated to additive Gaussian noise via a well chosen penalty.

This year, we achieved generalizations of these results beyond Gaussian denoising and characterized noise models for which the same phenomenon occurs. In particular, we proved that with (a variant of) Poisson noise and any prior probability on the unknowns, MMSE estimation can again be expressed as the solution of a penalized least-squares optimization problem. For additive scalar denoising, the phenomenon holds if and only if the noise distribution is log-concave, resulting in the perhaps surprising fact that scalar Laplacian denoising can be expressed as the solution of a penalized least-squares problem [22] Somewhere in the proofs appears an apparently new characterization of proximity operators of (nonconvex) penalties as subdifferentials of convex potentials [54].

7.1.5. Low-rank approximations: fast constrained algorithms

Participant: Jeremy Cohen.

Main collaborations: Nicolas Gillis (Univ. Mons, Belgium), Andersen Man Shun Ang (Univ. Mons, Belgium), Nicolas Nadisic (Univ. Mons, Belgium).

Low-Rank Approximations (LRA) aim at expressing the content of a multiway array by a sum of simpler separable arrays. Understood as a powerful unsupervised machine learning technique, LRA are most and foremost modern avatars of sparsity that are still not fully understood. In particular, algorithms to compute the parameters of LRA demend a lot of computer ressources and provide sub-optimal results. An important line of work over the last year has been to design efficient algorithms to compute constrained LRA, and in particular constrained low-rank tensor decompositions. This work has been carried out though a collaboration with the ERC project COLORAMAP of Nicolas Gillis (Univ. Mons, Belgium) and his PhD students Nicolas Nadisic (co-supervision) and Andersen Man Shun Ang.

Extrapolated Block-coordinate algorithms for fast tensor decompositions State-of-the-art algorithms for computing tensor decompositions are based on the idea that solving alternatively for smaller blocks of parameters is easier than solving the large problem at once. Despite showing nice convergence speeds, the obtained Block Coordinate Descent algorithms (BCD) are prone to being stuck near saddle points. We have shown in preliminary work, which is still ongoing, that BCD algorithms can be improved using Nesterov extrapolation in-between block updates. This improves empirical convergence speed in constrained and unconstrained tensor decompositions tremendously at almost no additional computation cost, and is therefore bound to have a large impact on the community [37].

Exact sparse nonnegative least-squares solutions to least-squares problems Another important LRA is Nonnegative Matrix factorization, which has found many diverse applications such as in remote sensing or automatic music transcription. Sometimes, imposing sparsity on parameters of NMF is crucial to be able to correctly process and interpret the output of NMF. However, sparse NMF has scarcely been studied, and its computation is challenging. In fact, even only a subproblem in a BCD approach, sparse nonnegative least-squares, is already NP-hard. We proposed to solve this sparse nonnegative least-squares problem exactly using a combinatorial algorithm. To reduce as much as possible the cost of solving this combinatorial problem, a Branch and Bound algorithm was proposed which, on average, reduces the computational complexity drastically. A next step will be to use this branch and bound algorithm as a brick for proposing an efficient algorithm for sparse NMF.

7.1.6. Algorithmic Exploration of Sparse Representations for Neurofeedback

Participant: Rémi Gribonval.

Claire Cury, Pierre Maurel & Christian Barillot (EMPENN Inria project-team, Rennes)

In the context of the HEMISFER (Hybrid Eeg-MrI and Simultaneous neuro-feedback for brain Rehabilitation) Comin Labs project (see Section 1), in collaboration with the EMPENN team, we validated a technique to estimate brain neuronal activity by combining EEG and fMRI modalities in a joint framework exploiting sparsity [82]. We then focused on directly estimating neuro-feedback scores rather than brain activity. Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) both allow measurement of brain activity for neuro-feedback (NF), respectively with high temporal resolution for EEG and high spatial resolution for fMRI. Using simultaneously fMRI and EEG for NF training is very promising to devise brain rehabilitation protocols, however performing NF-fMRI is costly, exhausting and time consuming, and cannot be repeated too many times for the same subject. We proposed a technique to predict NF scores from EEG recordings only, using a training phase where both EEG and fMRI NF are available [39]. A journal paper has been submitted.

7.2. Emerging activities on high-dimensional learning with neural networks

Participants: Rémi Gribonval, Himalaya Jain, Pierre Stock.

Main collaborations: Patrick Perez (Technicolor R & I, Rennes), Gitta Kutyniok (TU Berlin, Germany), Morten Nielsen (Aalborg University, Denmark), Felix Voigtlaender (KU Eichstätt, Germany), Herve Jegou and Benjamin Graham (FAIR, Paris)

dictionary learning, large-scale indexing, sparse deep networks, normalization, sinkhorn, regularization

Many of the data analysis and processing pipelines that have been carefully engineered by generations of mathematicians and practitioners can in fact be implemented as deep networks. Allowing the parameters of these networks to be automatically trained (or even randomized) allows to revisit certain classical constructions. Our team has started investigating the potential of such approaches both from an empirical perspective and from the point of view of approximation theory.

Learning compact representations for large-scale image search. The PhD thesis of Himalaya Jain [73], which received the Fondation Rennes 1 PhD prize this year, was dedicated to learning techniques for the design of new efficient methods for large-scale image search and indexing.

Equi-normalization of Neural Networks. Modern neural networks are over-parameterized. In particular, each rectified linear hidden unit can be modified by a multiplicative factor by adjusting input and output weights, without changing the rest of the network. Inspired by the Sinkhorn-Knopp algorithm, we introduced a fast iterative method for minimizing the l2 norm of the weights, equivalently the weight decay regularizer. It provably converges to a unique solution. Interleaving our algorithm with SGD during training improves the test accuracy. For small batches, our approach offers an alternative to batch- and group- normalization on CIFAR-10 and ImageNet with a ResNet-18. This work was presented at ICLR 2019 [41].

Approximation theory with deep networks. We study the expressivity of sparsely connected deep networks. Measuring a network's complexity by its number of connections with nonzero weights, or its number of neurons, we consider the class of functions which error of best approximation with networks of a given complexity decays at a certain rate. Using classical approximation theory, we showed that this class can be endowed with a norm that makes it a nice function space, called approximation space. We established that the presence of certain "skip connections" has no impact of the approximation space, and studied the role of the network's nonlinearity (also known as activation function) on the resulting spaces, as well as the benefits of depth. For the popular ReLU nonlinearity (as well as its powers), we related the newly identified spaces to classical Besov spaces, which have a long history as image models associated to sparse wavelet decompositions. The sharp embeddings that we established highlight how depth enables sparsely connected networks to approximate functions of increased "roughness" (decreased Besov smoothness) compared to shallow networks and wavelets. A preprint has been published and is under review for a journal [23].

7.3. Emerging activities on Nonlinear Inverse Problems

Compressive sensing, compressive learning, audio inpainting, phase estimation

7.3.1. Audio Inpainting and Denoising

Participants: Rémi Gribonval, Nancy Bertin, Clément Gaultier.

Main collaborations: Srdan Kitic (Orange, Rennes)

Inpainting is a particular kind of inverse problems that has been extensively addressed in the recent years in the field of image processing. Building upon our previous pioneering contributions [57], we proposed over the last five years a series of algorithms leveraging the competitive cosparse approach, which offers a very appealing trade-off between reconstruction performance and computational time, and its extensions to the incorporation of the so-called "social" into problems regularized by a cosparse prior. We exhibited a common framework allowing to tackle both denoising and declipping in a unified fashion [69]; these results, together with listening tests results that were specified and prepared in 2019 and will be run soon, will be included in an ongoing journal paper, to be submitted in 2020. This year, following Clément Gaultier Ph.D. defense [12], we progressed towards industrial transfer of these results through informal interaction with a company commercializing audio plugins, in particular with new developments to alleviate some artifacts absent from simulation but arising in real-world use cases.

7.4. Source Localization and Separation

Source separation, sparse representations, probabilistic model, source localization

Acoustic source localization is, in general, the problem of determining the spatial coordinates of one or several sound sources based on microphone recordings. This problem arises in many different fields (speech and sound enhancement, speech recognition, acoustic tomography, robotics, aeroacoustics...) and its resolution, beyond an interest in itself, can also be the key preamble to efficient source separation, which is the task of retrieving the source signals underlying a multichannel mixture signal. Over the last years, we proposed a general probabilistic framework for the joint exploitation of spatial and spectral cues [9], hereafter summarized as the "local Gaussian modeling", and we showed how it could be used to quickly design new models adapted to the data at hand and estimate its parameters via the EM algorithm. This model became the basis of a large number of works in the field, including our own. This accumulated progress lead, in 2015, to two main achievements: a new version of the Flexible Audio Source Separation Toolbox, fully reimplemented, was released [84] and we published an overview paper on recent and going research along the path of *guided* separation in a special issue of IEEE Signal Processing Magazine [11].

From there, our recent work divided into several tracks: maturity work on the concrete use of these tools and principles in real-world scenarios, in particular within the INVATE project and the collaboration with the startup 5^{th} dimension (see Sections 8.1.2 , 8.1.4), on the one hand; on the other hand, an emerging track on audio scene analysis with machine learning, evolved beyond the "localization and separation" paradigm, and is the subject of a more recent axis of research presented in Section 7.5 .

7.4.1. Towards Real-world Localization and Separation

Participants: Nancy Bertin, Frédéric Bimbot, Rémi Gribonval, Ewen Camberlein, Romain Lebarbenchon, Mohammed Hafsati.

Main collaborations: Emmanuel Vincent (MULTISPEECH Inria project-team, Nancy)

Based on the team's accumulated expertise and tools for localization and separation using the local Gaussian model, two real-world applications were addressed in the past year, which in turn gave rise to new research tracks.

First, our work within the voiceHome project (2015-2017), an OSEO-FUI industrial collaboration [0] aiming at developing natural language dialog in home applications, such as control of domotic and multimedia devices, in realistic and challenging situations (very noisy and reverberant environments, distant microphones) found its conclusion with the publication of a journal paper in a special issue of Speech Communication [14].

Accomplished progress and levers of improvements identified thanks to this project resulted in the granting of an Inria ADT (Action de Développement Technologique). This new development phase of the FASST software started in September 2017 and was achieved this year by the release of the third version of the toolbox, with significant progress towards efficient initialization, low latency and reduction of the computational burden.

In addition, evolutions of the MBSSLocate software initiated during this project led to a successful participation in the IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA) [77], and served as a baseline for the publication of the for the IEEE Signal Processing Cup 2019 [21]. The SP Cup was also fueled by the publicly available DREGON dataset 5 recorded in PANAMA, including noiseless speech and on-flight ego-noise recordings, devoted to source localization from a drone [117].

Finally, these progress also led to a new industrial transfer with the start-up 5th dimension (see Section 8.1.4). During this collaboration aiming at equipping a pair of glasses with an array of microphones and "smart" speech enhancement functionalities, we particularly investigated the impact of obstacles between microphones in the localization and separation performance, the selection of the best subset of microphones in the array for side speakers hidden by the head shadow, and the importance of speaker enrolment (learning spectral dictionaries of target users voices) in this use case.

7.4.2. Separation for Remixing Applications

Participants: Nancy Bertin, Rémi Gribonval, Mohammed Hafsati.

Main collaborations: Nicolas Epain (IRT b<>com, Rennes)

Second, through the Ph.D. of Mohammed Hafsati (in collaboration with the IRT b<>com with the INVATE project, see Section 8.1.2) started in November 2016, we investigated a new application of source separation to sound re-spatialization from Higher Order Ambisonics (HOA) signals [70], in the context of free navigation in 3D audiovisual contents. We studied the applicability conditions of the FASST framework to HOA signals and benchmarked localization and separation methods in this domain. Simulation results showed that separating sources in the HOA domain results in a 5 to 15 dB increase in signal-to-distortion ratio, compared to the microphone domain. These results were accepted for publication in the DAFx international conference [34]. We continued extending our methods following two tracks: hybrid acquisition scenarios, where the separation of HOA signals can be informed by complementary close-up microphonic signals, and the replacement of spectrogram NMF by neural networks for a better spectral adaptation of the models. Future work will include subjective evaluation of the developed workflows.

7.5. Towards comprehensive audio scene analysis

Source localization and separation, machine learning, room geometry, room properties, multichannel audio classification

[0] With partners: onMobile, Delta Dore, eSoftThings, Orange, Technicolor, LOUSTIC, Inria Nancy.

By contrast to the previous lines of work and results on source localization and separation, which are mostly focused on the *sources*, the following emerging activities consider the audio scene and its analysis in a wider sense, including the environment around the sources, and in particular the *room* they are included in, and their properties. This inclusive vision of the audio scene allows in return to revisit classical audio processing tasks, such as localization, separation or classification.

7.5.1. Room Properties: Estimating or Learning Early Echoes

Participants: Nancy Bertin, Diego Di Carlo, Clément Elvira.

Main collaborations: Antoine Deleforge (Inria Nancy – Grand Est), Ivan Dokmanic (University of Illinois at Urbana-Champaign, Coordinated Science Lab, USA), Robin Scheibler (Tokyo Metropolitan University, Tokyo, Japan), Helena Peic-Tukuljac (EPFL, Switzerland).

In [85] we showed that the knowledge of early echoes improved sound source separation performances, which motivates the development of (blind) echo estimation techniques. Echoes are also known to potentially be a key to the room geometry problem [65]. In 2019, two different approaches to this problem were explored.

As a competitive, yet similar approach to our previous work in [83], we proposed a new analytical method for off-the-grid early echoes estimation, based on continuous dictionaries and extensions of sparse recovery methods in this setting. From the well-known *cross-relation* between room impulse responses and signals in a "one source - two microphones" settings, the echo estimation problem can be recast as a Beurling-LASSO problem and solved with algorithms of this kind. This enables near-exact blind and off-grid echo retrieval from discrete-time measurements, and can outperform conventional methods by several orders of magnitude in precision, in an ideal case where the room impulse response is limited to a few weighted Diracs. Future work will include alternative initialization schemes, extensions to sparse-spectrum signals and noisy measurements, and applications to dereverberation and audio-based room shape reconstruction. This work, mostly lead by Clément Elvira, was submitted for publication in Icassp 2020.

On the other hand, the PhD thesis of Diego Di Carlo aims at applying the "Virtual Acoustic Space Traveler" (VAST) framework to the blind estimation of acoustic echoes, or other room properties (such as reverberation time, acoustic properties at the boundaries, etc.) Last year, we focused on identifying promising couples of inputs and outputs for such an approach, especially by leveraging the notions of relative transfer functions between microphones, the room impulse responses, the time-difference-of-arrivals, the angular spectra, and all their mutual relationships. In a simple yet common scenario of 2 microphones close to a reflective surface and one source (which may occur, for instance, when the sensors are placed on a table such as in voice-based assistant devices), we introduced the concept of microphone array augmentation with echoes (MIRAGE) and showed how estimation of early-echo characteristics with a learning-based approach is not only possible but can in fact benefit source localization. In particular, it allows to retrieve 2D direction of arrivals from 2 microphones only, an impossible task in anechoic settings. These first results were published in ICASSP [29]. In 2019, we improved the involved DNN architecture in MIRAGE and worked towards experimental validation of this result, by designing and recording a data set with annotated echoes in different conditions of reverberation. Future work will include extension of this data set, extension to more realistic and more complex scenarios (including more microphones, sources and reflective surfaces) and the estimation of other room properties such as the acoustic absorption at the boundaries, or ultimately, the room geometry. Some of these tracks currently benefit from the visit of Diego di Carlo to Bar-Ilan University (thanks to a MathSTIC doctoral outgoing mobility grant.)

7.5.2. Multichannel Audio Event and Room Classification

Participants: Marie-Anne Lacroix, Nancy Bertin.

Main collaborations: Pascal Scalart, Romuald Rocher (GRANIT Inria project-team, Lannion)

Typically, audio event detection and classification is tackled as a "pure" single-channel signal processing task. By constrast, audio source localization is the perfect example of multi-channel task "by construction". In parallel, the need to classify the type of scene or room has emerged, in particular from the rapid development of wearables, the "Internet of things" and their applications. The PhD of Marie-Anne Lacroix,

started in September 2018, combines these ideas with the aim of developing multi-channel, room-aware or spatially-aware audio classification algorithms for embedded devices. The PhD topic includes low-complexity and low-energy stakes, which will be more specifically tackled thanks to the GRANIT members area of expertise. During the first year of the PhD, we gathered existing data and identified the need for new simulations or recordings, and combined ideas from existing single-channel classification techniques with traditional spatial features in order to design several baseline algorithms for multi-channel joint localization and classification of audio events. The impact of feature quantization on classification performance is also currently under investigation and a participation to the 2020 edition of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) is envisioned.

7.6. Music Content Processing and Information Retrieval

Music structure, music language modeling, System & Contrast model, complexity

Current work developed in our research group in the domain of music content processing and information retrieval explore various information-theoretic frameworks for music structure analysis and description [58], in particular the System & Contrast model [1].

7.6.1. Modeling music by Polytopic Graphs of Latent Relations (PGLR)

Participants: Corentin Louboutin, Frédéric Bimbot.

The musical content observed at a given instant within a music segment obviously tends to share privileged relationships with its immediate past, hence the sequential perception of the music flow. But local music content also relates with distant events which have occurred in the longer term past, especially at instants which are metrically homologous (in previous bars, motifs, phrases, etc.) This is particularly evident in strongly "patterned" music, such as pop music, where recurrence and regularity play a central role in the design of cyclic musical repetitions, anticipations and surprises.

The web of musical elements can be described as a Polytopic Graph of Latent Relations (PGLR) which models relationships developing predominantly between homologous elements within the metrical grid.

For regular segments the PGLR lives on an n-dimensional cube(square, cube, tesseract, etc...), n being the number of scales considered simultaneously in the multiscale model. By extension, the PGLR can be generalized to a more or less regular n-dimensional polytopes.

Each vertex in the polytope corresponds to a low-scale musical element, each edge represents a relationship between two vertices and each face forms an elementary system of relationships.

The estimation of the PGLR structure of a musical segment can be obtained computationally as the joint estimation of the description of the polytope, the nesting configuration of the graph over the polytope (reflecting the flow of dependencies and interactions between the elements within the musical segment) and the set of relations between the nodes of the graph, with potentially multiple possibilities.

If musical elements are chords, relations can be inferred by minimal transport [79] defined as the shortest displacement of notes, in semitones, between a pair of chords. Other chord representations and relations are possible, as studied in [81] where the PGLR approach is presented conceptually and algorithmically, together with an extensive evaluation on a large set of chord sequences from the RWC Pop corpus (100 pop songs).

Specific graph configurations, called Primer Preserving Permutations (PPP) are extensively studied in [80] and are related to 6 main redundant sequences which can be viewed as canonical multiscale structural patterns.

In parallel, recent work has also been dedicated to modeling melodic and rhythmic motifs in order to extend the polytopic model to multiple musical dimensions.

Results obtained in this framework illustrate the efficiency of the proposed model in capturing structural information within musical data and support the view that musical content can be delinearised in order to better describe its structure. Extensive results are included in Corentin Louboutin's PhD [13], defended in March 2019 and which was awarded the Prix Jeune Chercheur Science et Musique, in October.

7.6.2. Exploring Structural Dependencies in Melodic Sequences using Neural Networks

Participants: Nathan Libermann, Frédéric Bimbot.

This work is carried out in the framework of a PhD, co-directed by Emmanuel Vincent (Inria-Nancy).

In order to be able to generate structured melodic phrases and section, we explore various schemes for modeling dependencies between notes within melodies, using deep learning frameworks.

A a first set of experiments, we have considered a GRU-based sequential learning model, studied under different learning scenarios in order to better understand the optimal architectures in this context that can achieve satisfactory results. By this means, we wish to explore different hypotheses relating to temporal non-invariance relationships between notes within a structural segment (motif, phrase, section).

We have defined three types of recursive architectures corresponding to different ways to exploit the local history of a musical note, in terms of information encoding and generalization capabilities.

Initially conducted on the Lakh MIDI dataset, experiments have switched to the Meertens Tune Collections data set (Dutch traditional melodies) and confirm the trends observed in [78], w.r.t. the utility of non-ergodic models for the generation of melodic segments.

Ongoing work is extending these findings to the design of specific NN architectures, which incorporate attention models, to account for this non-invariance of information across musical segments.

7.6.3. Graph Signal Processing for Multiscale Representations of Music Similarity

Participants: Valentin Gillot, Frédéric Bimbot.

"Music Similarity" is a multifaceted concept at the core of Music Information Retrieval (MIR). Among the wide range of possible definitions and approaches to this notion, a popular one is the computation of a so-called content-based similarity matrix (S), in which each coefficient is a similarity measure between descriptors of short time frames at different instants within a music piece or a collection of pieces.

Matrix S can be seen as the adjacency matrix of an underlying graph, embodying the local and non-local similarities between parts of the music material. Considering the nodes of this graph as a new set of indices for the original music frames or pieces opens the door to a "delinearized" representation of music, emphasizing its structure and its semiotic content.

Graph Signal Processing (GSP) is an emerging topic devoted to extend usual signal processing tools (Fourier analysis, filtering, denoising, compression, ...) to signals "living" on graphs rather than on the time line, and to exploit mathematical and algorithmic tools on usual graphs, in order to better represent and manipulate these signals. Toy applications of GSP concepts on music content in music resequencing and music inpainting are illustrating this trend.

From exploratory experiments, first observations point towards the following hypotheses :

- local and non-local structures of a piece are highlighted in the adjacency matrix built from a simple time-frequency representation of the piece,
- the first eigenvectors of the graph Laplacian provide a rough structural segmentation of the piece,
- clusters of frames built from the eigenvectors contain similar, repetitive sound sequences.

The goal of Valentin Gillot's PhD is to consolidate these hypotheses and investigate further the topic of Graph Signal Processing for music, with more powerful conceptual tools and experiments at a larger scale.

The core of the work will consist in designing a methodology and implement an evaluation framework so as to (i) compare different descriptors and similarity measures and their capacity to capture relevant structural information in music pieces or collection of pieces, (ii) explore the structure of musical pieces by refining the frame clustering process, in particular with a multi-resolution approach, (iii) identify salient characteristics of graphs in relation to mid-level structure models and (iv) perform statistics on the typical properties of the similarity graphs on a large corpus of music in relation to music genres and/or composers.

By the end of the PhD, we expect the release of a specific toolbox for music composition, remixing and repurposing using the concepts and algorithms developed during the PhD. First results obtained this year in music recomposition have proven very conclusive [32].

RAINBOW Project-Team

6. New Results

6.1. Optimal and Uncertainty-Aware Sensing

6.1.1. Tracking of Rigid Objects of Complex Shapes with a RDB-D Camera

Participants: Agniva Sengupta, Alexandre Krupa, Eric Marchand.

In the context of the iProcess project (see Section 8.3.8), we developed a method for accurately tracking the pose of rigid objects of complex shapes using a RGB-D camera [52]. This method only needs a coarse 3D geometric model of the object of interest represented as a 3D mesh. The tracking of the object is based on a joint minimization of geometric and photometric criteria and more particularly on a combination of point-to-plane distance minimization and photometric error minimization. The concept of successive "keyframes" was also used in this approach for minimizing possible drift of the tracking. The proposed approach was validated on both simulated and real data and the results experimentally demonstrated a better tracking accuracy than existing state-of-the-art 6-DoF object tracking methods, especially when dealing with low-textured objects, multiple coplanar faces, occlusions and partial specularities of the scene.

6.1.2. Deformable Object 3D Tracking based on Depth Information and Coarse Physical Model

Participants: Agniva Sengupta, Alexandre Krupa, Eric Marchand.

This research activity was also carried out in the context of the iProcess project (see Section 8.3.8) and will continue with the recent starting GentleMAN project (see Section 8.3.9). It focusses on the elaboration of approaches able to accurately track in real-time the deformation of soft objects using a RGB-D camera. The state-of-the-art approaches are currently relying on the use of Finite Element Model (FEM) to simulate the physics (mechanical behavior) of the deformable object. However, they suffer from the drawback of being excessively dependent on the accurate knowledge of the physical properties of the object being tracked (Young Modulus, Poisson's ratio, etc). This year, we proposed a first method that only required a coarse physical model of the object based on FEM whose parameters do not need to be precise [53]. The method consists in applying a set of virtual forces on the surface mesh of our coarse FEM model in such a way that it deforms to fit the current shape of the object. A point-to-plane distance error between the point cloud provided by the depth camera and the model mesh is iteratively minimized with respect to these virtual forces. The point of application of force is determined by an analysis of the error obtained from rigid tracking, which is done in parallel with the non-rigid tracking. The approach has been validated on simulated objects with ground-truth, as well on real objects of unknown physical properties and experimentally demonstrated that accurate tracking of deformable objects can be achieved without the need of a precise physical model.

6.1.3. Trajectory Generation for Optimal State Estimation

Participants: Marco Cognetti, Paolo Robuffo Giordano.

This activity addresses the general problem of *active sensing* where the goal is to analyze and synthesize optimal trajectories for a robotic system that can maximize the amount of information gathered by the (few) noisy outputs (i.e., sensor readings) while at the same time reducing the negative effects of the process/actuation noise. Over the last years we have developed a general framework for solving *online* the active sensing problem by continuously replanning an optimal trajectory that maximizes a suitable norm of the Constructibility Gramian (CG), while also coping with a number of constraints including limited energy and feasibility. The results obtained so far have been generalized and summarized in [27], where the online trajectory replanning for CG maximization has been applied to two relevant case studies (unicycle and quadrotor) and validated via a large statistical campaign. We are actually working towards the extension of this machinery to the case of realization of a robot task (e.g., reaching and grasping for a mobile manipulator), and to the mutual localization problem for a multi-robot group.

6.1.4. Robotic manipulators in Physical Interaction with the Environment

Participant: Claudio Pacchierotti.

As robotic systems become more flexible and intelligent, they must be able to move into environments with a high degree of uncertainty or clutter, such as our homes, workplaces, and the outdoors. In these unstructured scenarios, it is possible that the body of the robot collides with its surroundings. As such, it would be desirable to characterise these contacts in terms of their location and interaction forces. We worked to address the problem of detecting and isolating collisions between a robotic manipulator and its environment, using only on-board joint torque and position sensing [37]. We presented an algorithm based on a particle filter that, under some assumptions, is able to identify the contact location anywhere on the robot body. It requires the robot to perform small exploratory movements, progressively integrating the new sensing information through a Bayesian framework. The method assumes negligible friction forces, convex contact surfaces, and linear contact stiffness. Compared to existing approaches, it allows this detection to be carried in almost all the surface of the robot's body. We tested the proposed approach both in simulation and in a real environment. Experiments in simulation showed that our approach outperformed two other methods that made simpler assumptions. Experiments in a real environment using a robot with joint torque sensors showed the applicability of the method to real world scenarios and its ability to cope with situations where the algorithm's assumptions did not hold.

6.1.5. Cooperative Localization using Interval Analysis

Participants: Ide Flore Kenmogne Fokam, Vincent Drevelle, Eric Marchand.

In the context of multi-robot fleets, cooperative localization consists in gaining better position estimate through measurements and data exchange with neighboring robots. Positioning integrity (i.e., providing reliable position uncertainty information) is also a key point for mission-critical tasks, like collision avoidance. The goal of this work is to compute position uncertainty volumes for each robot of the fleet, using a decentralized method (i.e., using only local communication with the neighbors). The problem is addressed in a bounded-error framework, with interval analysis and constraint propagation methods. These methods enable to provide guaranteed position error bounds, assuming bounded-error measurements. They are not affected by over-convergence due to data incest, which makes them a well sound framework for decentralized estimation. Quantifier elimination techniques have been used to consider uncertainty in the landmarks positions without adding pessimism in the computed solution. This work has been applied to cooperative localization of UAVs, based on image and range measurements [20].

6.2. Advanced Sensor-Based Control

6.2.1. Sensor-based Trajectory Planning for quadrotor UAVs

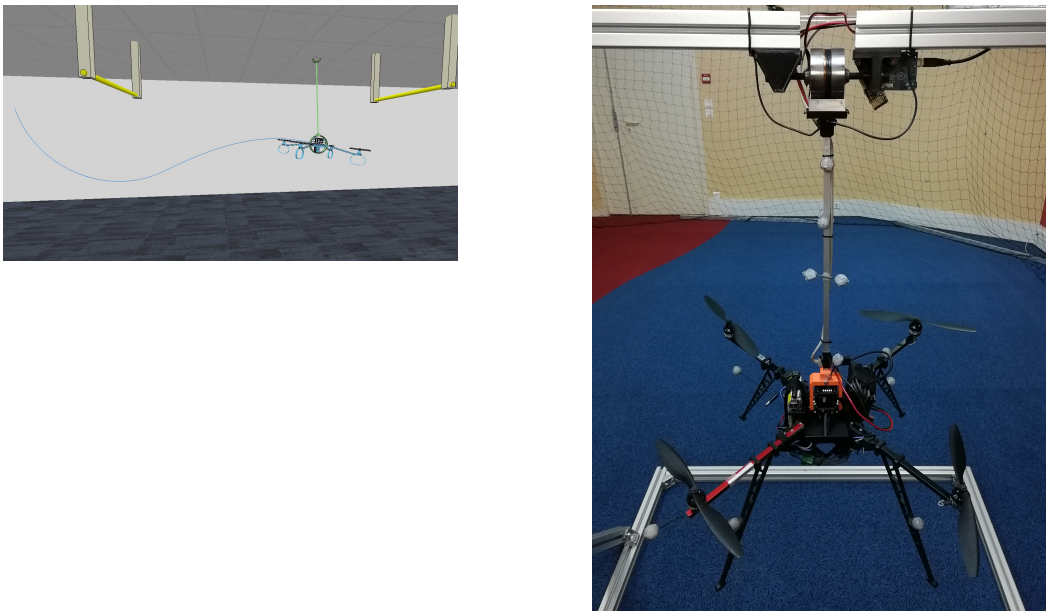
Participants: François Chaumette, Paolo Robuffo Giordano.

In the context of developing robust navigation strategies for quadrotor UAVs with onboard cameras and IMUs, we considered the problem of planning minimum-time trajectors in a cluttered environment for reaching a goal while coping with actuation and sensing constraints [25]. In particular, we considered a realistic model for the onboard camera that considers limited fov and possible occlusions due to obstructed visibility (e.g., presence of obstacles). Whenever the camera can detect landmarks in the environment, the visual cues can be used to drive a state estimation algorithm (a EKF) for updating the current estimation of the UAV state (its pose and velocity). However, beacause of the sensing constraints, the possibility of detecting and tracking the landmarks may be lost while moving in the environment. Therefore, we proposed a robust "perception-aware" planning strategy, based on the bi-directional A^* planner,

6.2.2. UAVs in Physical Interaction with the Environment

Participants: Quentin Delamare, Paolo Robuffo Giordano.

Most research in UAVs deals with either contact-free cases (the UAVs must avoid any contact with the environment), or "static" contact cases (the UAVs need to exert some forces on the environment in quasi-static conditions, reminiscent of what has been done with manipulator arms). Inspired by the vast literature on robot locomotion (from, e.g., the humanoid community), in this research topic we aim at exploiting the contact with the environment for helping a UAV maneuvering in the environment, in the same spirit in which we humans (and, supposedly, humanoid robots) use our legs and arms when navigating in cluttered environments for helping in keeping balance, or perform maneuvers that would be, otherwise, impossible. During last year we have considered the modeling, control and trajectory planning problem for a planar UAV equipped with a 1 DoF actuated arm capable of hooking at some pivots in the environment. This UAV (named MonkeyRotor) needs to "jump" from one pivot to the next one by exploiting the forces exchanged with the environment (the pivot) and its own actuation system (the propellers), see Fig. 8 (a). We are currently finalizing a real prototype (Fig. 8 (b)) for obtaining an experimental validation of the whole approach [1].



(a) (b)

Figure 8. UAVs in Physical Interaction with the Environment. a) The simulated MonkeyRotor performing a hook-to-hook maneuver. b) The prototype currently under finalization.

6.2.3. Trajectory Generation for Minimum Closed-Loop State Sensitivity

Participants: Pascal Brault, Quentin Delamare, Paolo Robuffo Giordano.

The goal of this research activity is to propose a new point of view in addressing the control of robots under parametric uncertainties: rather than striving to design a sophisticated controller with some robustness guarantees for a specific system, we propose to attain robustness (for any choice of the control action) by suitably shaping the reference motion trajectory so as to minimize the *state sensitivity* to parameter uncertainty of the resulting closed-loop system. During this year, we have extended the existing minimization framework to also include the notion of "input sensitivity", which allows to obtain trajectories whose realization (in perturbed conditions) leaves the control inputs unchanged to the largest extent. Such a feature is relevant whenever dealing with, e.g., limited actuation since it guarantees that, even under model perturbations, the inputs do not deviate too much from their nominal values. This novel input sensitivity has been combined

with the previously introduced notion of state sensitivity and validated both via monte-carlo simulations and experimentally with a unicycle robot in a large number of tests [1].

6.2.4. Visual Servoing for Steering Simulation Agents

Participants: Axel Lopez Gandia, Eric Marchand, François Chaumette, Julien Pettré.

This research activity is dedicated to the simulation of human locomotion, and more especially to the simulation of the visuomotor loop that controls human locomotion in interaction with the static and moving obstacles of its environment. Our approach is based on the principles of visual servoing for robots. To simulate visual perception, an agent perceives its environment through a virtual camera located in the position of its head. The visual input is processed by each agent in order to extract the relevant information for controlling its motion. In particular, the optical flow is computed to give the agent access to the relative motion of visible objects around it. Some features of the optical flow are finally computed to estimate the risk of collision with obstacle. We have established the mathematical relations between those visual features and the agent's self motion. Therefore, when necessary, the agent motion is controlled and adjusted so as to cancel the visual features indicating a risk of future collision [22], [46].

6.2.5. Strategies for Crowd Simulation Agents

Participants: Wouter Van Toll, Julien Pettré.

This research activity is dedicated to the simulation of crowds based on microscopic approaches. In such approaches, agents move according to local models of interactions that give them the capacity to adjust to the motion of neighbor agents. These purely local rules are not sufficient to produce high-quality long term trajectories through their environment. We provide agents with the capacity to establish mid-term strategies to move through their environment, by establishing a local plan based on their prediction of their surroundings and by verifying regularly this prediction remains valid. In the case validity is not checked, planning a new strategy is triggered [55].

6.2.6. Study of human locomotion to improve robot navigation

Participants: Florian Berton, Julien Bruneau, Julien Pettré.

This research activity is dedicated to the study of human gaze behaviour during locomotion. This activity is directly linked to the previous one on simulation, as human locomotion study results will serve as an input for the design of novel models for simulation. We are interested in the study of the activity of the gaze during locomotion that, in addition to the classical study of kinematics motion parameters, provides information on the nature of visual information acquired by humans to move, and the relative importance of visual elements in their surroundings [36].

6.2.7. Robot-Human Interactions during Locomotion

Participants: Javad Amirian, Fabien Grzeskowiak, Marie Babel, Julien Pettré.

This research activity is dedicated to the design of robot navigation techniques to make them capable of safely moving through a crowd of people. We are following two main research paths. The first one is dedicated to the prediction of crowd motion based on the state of the crowd as sensed by a robot. The second one is dedicated to the creation of a virtual reality platform that enables robots and humans to share a common virtual space where robot control techniques can be tested with no physical risk of harming people, as they remain separated in the physical space. This year, we have delivered techniques for the short term prediction of human locomotion trajectories [34], [35] and robot-human collision avoidance [39].

6.2.8. Visual Servoing for Cable-Driven Parallel Robots

Participant: François Chaumette.

This study is done in collaboration with IRT Jules Verne (Zane Zake, Nicolo Pedemonte) and LS2N (Stéphane Caro) in Nantes (see Section 7.2.2). It is devoted to the analysis of the robustness of visual servoing to modeling and calibration errors for cable-driven parallel robots. The modeling of the closed loop system has been derived, from which a Lyapunov-based stability analysis allowed exhibiting sufficient conditions for ensuring its stability. Experimental results have validated the theoretical results obtained and shown the high robustness of visual servoing for this sort of robots [30], [56].

6.2.9. Visual Exploration of an Indoor Environment

Participants: Benoît Antoniotti, Eric Marchand, François Chaumette.

This study is done in collaboration with the Creative company in Rennes (see Section 6.2.9). It is devoted to the exploration of indoor environments by a mobile robot, Pepper typically (see Section 5.4.2) for a complete and accurate reconstruction of the environment. The exploration strategy we are currently developing is based on maximizing the entropy generated by a robot motion.

6.2.10. Deformation Servoing of Soft Objects

Participant: Alexandre Krupa.

Nowadays robots are mostly used to manipulate rigid objects. Manipulating deformable objects remains challenging due to the difficulty of accurately predicting the object deformations. This year, we developed a model-free deformation servoing method able to do an online estimation of the deformation Jacobian that relates the motion of the robot end-effector to the deformation of a manipulated soft object. The first experimental results are encouraging since they showed that our model-free visual servoing approach based on online estimation provides similar results than a model-based approach based on physics simulation that requires accurate knowledge of the physical properties of the object to deform. This approach has been recently submitted to the ICRA'20 conference.

6.2.11. Multi-Robot Formation Control

Participant: Paolo Robuffo Giordano.

Most multi-robot applications must rely on relative sensing among the robot pairs (rather than absolute/external sensing such as, e.g., GPS). For these systems, the concept of rigidity provides the correct framework for defining an appropriate sensing and communication topology architecture. In several previous works we have addressed the problem of coordinating a team of quadrotor UAVs equipped with onboard sensors (such as distance sensors or cameras) for cooperative localization and formation control under the rigidity framework. In [9] an interesting interplay between the rigidity formalism and notions of parallel robotics has been studied, showing how well-known tools from the parallel robotics community can be applied to the multi-robot case, and how these tools can be used for characterizing the stability and singularities of the typical formation control/localization algorithms.

In [17], the problem of distributed leader selection has been addressed by considering agents with a second-order dynamics, thus closer to physical robots that have some unavoidable inertia when moving. This work has extended a previous strategy developed for a first-order case and ported it to the second-order: the proposed algorithm is able to periodically select at runtime the 'best' leader (among the neighbors of the current leader) for maximizing the tracking performance of an external trajectory reference while maintaining a desired formation for the group. The approach has been validated via numerical simulations.

6.2.12. Coupling Force and Vision for Controlling Robot Manipulators

Participants: Alexander Oliva, François Chaumette, Paolo Robuffo Giordano.

The goal of this activity is about coupling visual and force information for advanced manipulation tasks. To this end, we plan to exploit the recently acquired Panda robot (see Sect. 5.4.4), a state-of-the-art 7-dof manipulator arm with torque sensing in the joints, and the possibility to command torques at the joints or forces at the end-effector. Thanks to this new robot, we plan to study how to optimally combine the torque sensing and control strategies that have been developed over the years to also include in the loop the feedback from a vision sensor (a camera). In fact, the use of vision in torque-controlled robot is quite limited because of many issues, among which the difficulty of fusing low-rate images (about 30 Hz) with high-rate torque commands (about 1 kHz), the delays caused by any image processing and tracking algorithms, and the unavoidable occlusions that arise when the end-effector needs to approach an object to be grasped.

Towards this goal, this year we have considered the problem of identification of the dynamical model for the Panda robot [18], by suitably exploiting tools from identification theory. The identified model has been validated in numerous tests on the real robot with very good results and accuracy. A special feature of the model is the inclusion of a (realistic) friction term that accounts well for joint friction (a term that is usually neglected in dynamical model identification).

6.2.13. Subspace-based visual servoing

Participant: Eric Marchand.

To date most of visual servoing approaches have relied on geometric features that have to be tracked and matched in the image. Recent works have highlighted the importance of taking into account the photometric information of the entire images. This leads to direct visual servoing (DVS) approaches. The main disadvantage of DVS is its small convergence domain compared to conventional techniques, which is due to the high non-linearities of the cost function to be minimized. We proposed to project the image on an orthogonal basis (PCA) and then servo on either images reconstructed from this new compact set of coordinates or directly on these coordinates used as visual features [23]. In both cases we derived the analytical formulation of the interaction matrix. We show that these approaches feature a better behavior than the classical photometric visual servoing scheme allowing larger displacements and a satisfactory decrease of the error norm thanks to a well modelled interaction matrix.

6.2.14. Wheelchair Autonomous Navigation for Fall Prevention

Participants: Solenne Fortun, Marie Babel.

The Prisme project (see Section 8.1.4) is devoted to fall prevention and detection of inpatients with disabilities. For wheelchair users, falls typically occur during transfer between the bed and the wheelchair and are mainly due to a bad positioning of the wheelchair. In this context, the Prisme project addresses both fall prevention and detection issues by means of a collaborative sensing framework. Ultrasonic sensors are embedded onto both a robotized wheelchair and a medical bed. The measured signals are used to detect fall and to automatically drive the wheelchair near the bed at an optimal position determined by occupational therapists. This year, we finalized the related control framework based on sensor-based servoing principles. We validated the proposed solution through usage tests within the Rehabilitation Center of Pôle Saint Hélier (Rennes).

6.3. Haptic Cueing for Robotic Applications

6.3.1. Wearable Haptics

Participants: Marco Aggravi, Claudio Pacchierotti.

We worked on developing a novel modular wearable finger interface for cutaneous and kinesthetic interaction [11], shown in Fig. 9 . It is composed of a 3-DoF fingertip cutaneous device and a 1-DoF finger kinesthetic exoskeleton, which can be either used together as a single device or separately as two different devices. The 3-DoF fingertip device is composed of a static body and a mobile platform. The mobile platform is capable of making and breaking contact with the finger pulp and re-angle to replicate contacts with arbitrarily oriented surfaces.

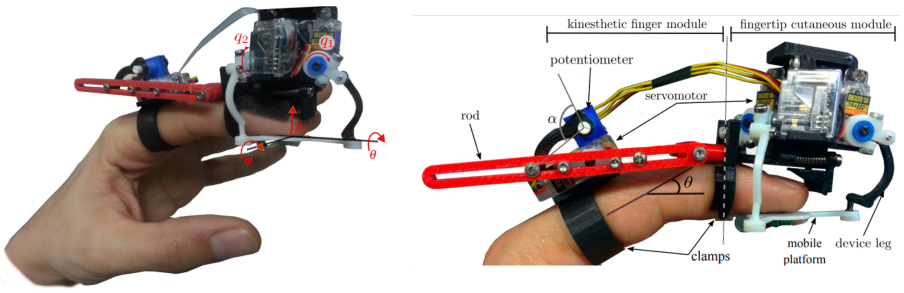


Figure 9. The proposed wearable device. It can provide both cutaneous feedback at the fingertip and kinesthetic feedback at the finger.

The 1-DoF finger exoskeleton provides kinesthetic force to the proximal and distal interphalangeal finger articulations using one servo motor grounded on the proximal phalanx. Together with the wearable device, we designed three different position, force, and compliance control schemes. We also carried out three human subjects experiments, enrolling a total of 40 different participants: the first experiment considered a curvature discrimination task, the second one a robotassisted palpation task, and the third one an immersive experience in Virtual Reality. Results showed that providing cutaneous and kinesthetic feedback through our device significantly improved the performance of all the considered tasks. Moreover, although cutaneous-only feedback showed promising performance, adding kinesthetic feedback improved most metrics. Finally, subjects ranked our device as highly wearable, comfortable, and effective.

On the same line of research, this year we guest edited a Special Issue on the IEEE Transactions on Haptics [26]. Thirteen papers on the topic have been published.

6.3.2. Mid-Air Haptic Feedback

Participants: Claudio Pacchierotti, Thomas Howard.

GUIs have been the gold standard for more than 25 years. However, they only support interaction with digital information indirectly (typically using a mouse or pen) and input and output are always separated. Furthermore, GUIs do not leverage our innate human abilities to manipulate and reason with 3D objects. Recently, 3D interfaces and VR headsets use physical objects as surrogates for tangible information, offering limited malleability and haptic feedback (e.g., rumble effects). In the framework of project H-Reality (Sect. 8.3.5), we are working to develop novel mid-air haptics paradigm that can convey the information spectrum of touch sensations in the real world, motivating the need to develop new, natural interaction techniques.

In this respect, we started working on investigating the recognition of local shapes using mid-air ultrasound haptics [45]. We have presented a series of human subject experiments investigating important perceptual aspects related to the rendering of 2D shapes by an ultrasound haptic interface (the Ultrahaptics STRATOS platform). We carried out four user studies aiming at evaluating (i) the absolute detection threshold for a static focal point rendered via amplitude modulation, (ii) the absolute detection and identification thresholds for line patterns rendered via spatiotemporal modulation, (iii) the ability to discriminate different line orientations, and (iv) the ability to perceive virtual bumps and holes.

Our results show that focal point detection thresholds are situated around 560Pa peak acoustic radiation pressure, with no evidence of effects of hand movement on detection. Line patterns rendered through spatiotemporal modulation were detectable at lower pressures, however their shape was generally not recognized as a line below a similar threshold of approx. 540Pa peak acoustic radiation pressure. We did not find any significant effect of line orientation relative to the hand both in terms of detection thresholds and in terms of correct identification of line orientation.

6.3.3. Tangible objects in VR and AR

Participant: Claudio Pacchierotti.

Still in the framework of the H-Reality project (Sect. 8.3.5), we studied the role of employing simple tangible objects in VR and AR scenarios, to improve the illusion of telepresence in these environments. We started by investigating the role of haptic sensations when interacting with tangible objects. Tangible objects are used in Virtual Reality to provide human users with distributed haptic sensations when grasping virtual objects. To achieve a compelling illusion, there should be a good correspondence between the haptic features of the tangible object and those of the corresponding virtual one, i.e., what users see in the virtual environment should match as much as possible what they touch in the real world. For this reason, we aimed at quantifying how similar tangible and virtual objects need to be, in terms of haptic perception, to still feel the same [40]. As it is often not possible to create tangible replicas of all the virtual objects in the scene, it is indeed important to understand how different tangible and virtual objects can be without the user noticing. Of course, the visuohaptic perception of objects encompasses several different dimensions, including the object's size, shape, mass, texture, and temperature. We started by addressing three representative haptic features - width, local orientation, and curvature, - which are particularly relevant for grasping. We evaluated the just-noticeable difference (JND) when grasping, with a thumb-index pinch, a tangible object which differ from a seen virtual one on the above three important haptic features. Results show JND values of 5.75%, 43.8%, and 66.66% of the reference shape for the width, local orientation, and local curvature features, respectively.

As we mentioned above, for achieving a compelling illusion during interaction in VR, there should be a good correspondence between what users see in the virtual environment and what they touch in the real world. The haptic features of the tangible object should – up to a certain extent – match those of the corresponding virtual one. We worked on an innovative approach enabling the use of few tangible objects to render many virtual ones [41]. Toward this objective, we present an algorithm which analyses different tangible and virtual objects to find the grasping strategy best matching the resultant haptic pinching sensation. Starting from the meshes of the considered objects, the algorithm guides users towards the grasping pose which best matches what they see in the virtual scene with what they feel when touching the tangible object. By selecting different grasping positions according to the virtual object to render, it is possible to use few tangible objects to render multiple virtual ones. We tested our approach in a user study. Twelve participants were asked to grasp different virtual objects, all rendered by the same tangible one. For every virtual object, our algorithm found the best pinching match on the tangible one, and guided the participant toward that grasp. Results show that our algorithm was able to well combine several haptically-salient object features to find corresponding pinches between the given tangible and virtual objects. At the end of the experiment, participants were also asked to guess how many tangible objects were used during the experiment. No one guessed that we used only one, proof of a convincing experience.

6.3.4. Wearable haptics for an Augmented Wheelchair Driving Experience

Participants: Louise Devigne, François Pasteau, Marco Aggravi, Claudio Pacchierotti, Marie Babel.

Smart powered wheelchairs can increase mobility and independence for people with disability by providing navigation support. For rehabilitation or learning purposes, it would be of great benefit for wheelchair users to have a better understanding of the surrounding environment while driving. Therefore, a way of providing navigation support is to communicate information through a dedicated and adapted feedback interface.

We then envisaged the use of wearable vibrotactile haptics, i.e. two haptic armbands, each composed of four evenly-spaced vibrotactile actuators. With respect to other available solutions, our approach provides rich navigation information while always leaving the patient in control of the wheelchair motion. We then conducted experiments with volunteers who experienced wheelchair driving in conjunction with the use of the armbands to provide drivers with information either on the presence of obstacles. Results show that providing information on closest obstacle position improved significantly the safety of the driving task (least number of collisions). This work is jointly conducted in the context of ADAPT project (Sect. 8.3.6) and ISI4NAVE associate team (Sect. 8.4.1.1).

6.4. Shared Control Architectures

6.4.1. Shared Control for Remote Manipulation

Participants: Firas Abi Farraj, Paolo Robuffo Giordano, Claudio Pacchierotti, Rahaf Rahal.

As teleoperation systems become more sophisticated and flexible, the environments and applications where they can be employed become less structured and predictable. This desirable evolution toward more challenging robotic tasks requires an increasing degree of training, skills, and concentration from the human operator. For this reason, researchers started to devise innovative approaches to make the control of such systems more effective and intuitive. In this respect, shared control algorithms have been investigated as one the main tools to design complex but intuitive robotic teleoperation system, helping operators in carrying out several increasingly difficult robotic applications, such as assisted vehicle navigation, surgical robotics, brain-computer interface manipulation, rehabilitation. This approach makes it possible to share the available degrees of freedom of the robotic system between the operator and an autonomous controller. The human operator is in charge of imparting high level, intuitive goals to the robotic system; while the autonomous controller translates them into inputs the robotic system can understand. How to implement such division of roles between the human operator and the autonomous controller highly depends on the task, robotic system, and application. Haptic feedback and guidance have been shown to play a significant and promising role in shared control applications. For example, haptic cues can provide the user with information about what the autonomous controller is doing or is planning to do; or haptic force can be used to gradually limit the degrees of freedom available to the human operator, according to the difficulty of the task or the experience of the user. The dynamic nature of haptic guidance enables us to design very flexible robotic system, which can easily and rapidly change the division of roles between the user and autonomous controller.

Along this general line of research, during this year we gave the following contributions:

- in [51] we proposed a shared control algorithm for remote telemanipulation of redundant robots able to fuse the task-prioritized control architecture (for handling the concurrent realization of multiple tasks) with haptic guidance techniques. In particular, we developed a suitable passivity-preserving strategy based on energy tanks for always guaranteeing stability despite the possible presence of autonomous tasks that could generate an increase of energy during operation. The approach has been validated with extensive simulative results in a realistic environment.
- in [6] we have considered a shared control algorithm for telemanipulation that embeds the presence of a grasping planner for guiding the operator towards suitable grasping poses. The operator retains control of the end-effector motion and eventual grasping location, but she/he is assisted by the autonomy (via force cues) in navigating towards good grasps, as classified by the grasping planner that takes as input a RGBD image of the scene and computes a set of grasping poses along the object contour.
- in [50] we have presented two haptic shared-control approaches for robotic cutting. They are designed to assist the human operator by enforcing different nonholonomic-like constraints representative of the cutting kinematics. To validate this approach, we carried out a human-subject experiment in a real cutting scenario. We compared our shared-control techniques with each other and with a standard haptic teleoperation scheme. Results show the usefulness of assisted control schemes in complex applications such as cutting.

6.4.2. Teleoperation of Flexible Needle with Haptic Feedback and Ultrasound Guidance

Participants: Jason Chevrie, Alexandre Krupa, Marie Babel.

Needle insertion procedures under ultrasound guidance are commonly used for diagnosis and therapy. This kind of intervention can greatly benefit from robotic systems to improve their accuracy and success rate. In the past years, we have developed a robotic framework dedicated to 3D steering of beveled-tip flexible needle in order to autonomously reach a desired target in the tissues by ultrasound visual servoing using a 3D ultrasound probe. This year we have proposed a real-time semi-automatic teleoperation framework that enables the user to directly control the trajectory of the needle tip during its insertion via a haptic interface [38]. The framework

enables the user to intuitively guide the trajectory of the needle tip in the ultrasound 3D volume while the controller handles the complexity of the 6D motion that needs to be applied to the needle base. A mean targeting accuracy of 2.5 mm has been achieved in gelatin phantoms and different ways to provide the haptic feedback as well as different levels of control given to the user on the tip trajectory have been compared. Limiting the user input to the insertion speed while automatically controlling the trajectory of the needle tip seems to provide a safer insertion process, however it may be too constraining and can not handle situations where more control over the tip trajectory is required, for example if unpredicted obstacles need to be avoided. On the contrary, giving the full control of the 3D tip velocity to the user and applying a haptic feedback to guide the user toward the target proved to maintain a low level of needle bending and tissue deformation.

6.4.3. Needle Comanipulation with Haptic Guidance

Participants: Hadrien Gurnel, Alexandre Krupa.

The objective of this work is to provide assistance during manual needle steering for biopsies or therapy purposes (see Section 7.2.3). At the difference of our work presented in Section 6.4.2 where a robotic system is used to steer the needle, we propose in this study another way of assistance where the needle is collaboratively manipulated by the physician and a haptic device. The principle of our approach is to provide haptic cue feedback to the clinician in order to help him during his manual gesture [43]. We elaborated 5 different haptic-guidance strategies to assist the needle pre-positioning and pre-orienting on a pre-defined insertion point, and with a pre-planned desired incidence angle. The haptic guides rely on the position and orientation errors between the needle, the entry point and the desired angle of incidence toward the target, which are computed from the measurements provided by an electromagnetic tracker. Each of the guide implements a different Guiding Virtual Fixture producing haptic cues that attract the needle towards a point or a trajectory in space with different force feedback applied on the user's hand manipulating the needle. A two-step evaluation was conducted to assess the performance and ergonomy of each haptic guide, and compare them to the unassisted reference gesture. The first evaluation stage [44] involved two physicians both experts in needle manipulation at Rennes University Hospital. The performance results showed that, compared to the unassisted gesture, the positioning accuracy was enhanced with haptic guidance. The second evaluation stage [43] was a user study with twelve participants. From this second study it results that the most constraining guide allows to perform the gesture with the best accuracy, lower time duration and highest level of ergonomy.

6.4.4. Shared Control of a Wheelchair for Navigation Assistance

Participants: Louise Devigne, Marie Babel.

Power wheelchairs allow people with motor disabilities to have more mobility and independence. However, driving safely such a vehicle is a daily challenge particularly in urban environments while navigating on sidewalks, negotiating curbs or dealing with uneven grounds. Indeed, differences of elevation have been reported to be one of the most challenging environmental barrier to negotiate, with tipping and falling being the most common accidents power wheelchair users encounter. It is thus our challenge to design assistive solutions for power wheelchair navigation in order to improve safety while navigating in such environments. To this aim, we proposed a shared-control algorithm which provides assistance while navigating with a wheelchair in an environment consisting of negative obstacles. We designed a dedicated sensor-based control law allowing trajectory correction while approaching negative obstacles e.g. steps, curbs, descending slopes. This shared control method takes into account the humanin-the loop factor. In this study, our solution the ability of our system to ensure a safe trajectory while navigating on a sidewalk is demonstrated through simulation, thus providing a proof-of-concept of our method [42].

6.4.5. Wheelchair-Human Interactions during crossing situations

Participants: Marie Babel, Julien Pettré.

Designing smart powered wheelchairs requires to better understand interactions between walkers and such vehicles. We focus on collision avoidance task between a power wheelchair (fully operated by a human) and a walker, where the difference in the nature of the agents (weight, maximal speed, acceleration profiles) results into asymmetrical physical risk in case of a collision, for example due to the protection power wheelchair provides to its driver, or the higher energy transferred to the walker during head-on collision.

We then conducted experiments with Results show that walkers set more conservative strategies when interacting with a power wheelchair. These results can then be linked to the difference in the physical characteristics of the walkers and power wheelchairs where asymmetry in the physical risks raised by collisions influence the strategies performed by the walkers in comparison with a similar walker-walker situation. This gives interesting insights in the task of modeling such interactions, indicating that geometrical terms are not sufficient to explain behaviours, physical terms linked to collision momentum should also be considered [49][62].

6.4.6. Multisensory power wheelchair simulator

Participants: Guillaume Vailland, Louise Devigne, François Pasteau, Marie Babel.

Power wheelchair driving is a challenging task which requires good visual, cognitive and visuo-spatial abilities. Besides, a power wheelchair can cause material damage or represent a danger of injury for others or oneself if not operated safely. Therefore, training and repeated practice are mandatory to acquire safe driving skills to obtain power wheelchair prescription from therapists. However, conventional training programs may reveal themselves insufficient for some people with severe impairments. In this context, Virtual Reality offers the opportunity to design innovative learning and training programs while providing realistic wheelchair driving experience within a virtual environment. We then proposed a user-centered design of a multisensory power wheelchair simulator [59][58]. This simulator addresses classical virtual experience drawbacks such as cybersickness and sense of presence by combining 3D visual rendering, haptic and vestibular feedback. It relies on a modular and versatile workflow enabling not only easy interfacing with any virtual display, but also with any user interface such as wheelchair controllers or feedback devices. First experiments with able-bodied people shown that vestibular feedback activation increases the Sense of Presence and decreases cybersickness [54].

SIROCCO Project-Team

7. New Results

7.1. Visual Data Analysis

Scene depth, Scene flows, 3D modeling, Light-fields, 3D point clouds

7.1.1. Scene depth estimation from light fields

Participants: Christine Guillemot, Xiaoran Jiang, Jinglei Shi.

While there exist scene depth estimation methods, these methods, mostly designed for stereo content or for pairs of rectified views, do not effectively apply to new imaging modalities such as light fields. We have focused on the problem of *scene depth estimation* for every viewpoint of a dense light field, exploiting information from only a sparse set of views [24]. This problem is particularly relevant for applications such as light field reconstruction from a subset of views, for view synthesis, for 3D modeling and for compression. Unlike most existing methods, the proposed algorithm computes disparity (or equivalently depth) for every viewpoint taking into account occlusions. In addition, it preserves the continuity of the depth space and does not require prior knowledge on the depth range.

We have then proposed a learning based depth estimation framework suitable for both densely and sparsely sampled light fields. The proposed framework consists of three processing steps: initial depth estimation, efficient fusion with occlusion handling and refinement. The estimation can be performed from a flexible subset of input views. The fusion of initial disparity estimates, relying on two warping errors measures, allows us to have an accurate estimation in occluded regions and along the contours. The use of trained neural networks has the advantage of a limited computational cost at estimation time. In contrast with methods relying on the computation of cost volumes, the proposed approach does not need any prior information on the disparity range. Experimental results show that the proposed method outperforms state-of-the-art light fields depth estimation methods for a large range of baselines [15].

The training of the proposed neural networks based architecture requires having ground truth disparity (or depth) maps. Although a few synthetic datasets exist for dense light fields with ground truth depth maps, no such dataset exists for sparse light fields with large baselines. This lack of training data with ground truth depth maps is a crucial issue for supervised learning of neural networks for depth estimation. We therefore created two datasets, namely SLFD and DLFD, containing respectively sparsely sampled and densely sampled synthetic light fields. To our knowledge, SLFD is the first available dataset providing sparse light field views and their corresponding ground truth depth and disparity maps. The created datasets have been made publicly available together with the code and the trained models.

7.1.2. Scene flow estimation from light fields

Participants: Pierre David, Christine Guillemot.

We have addressed the problem of scene flow estimation from sparsely sampled video light fields. Scene flows can be seen as 3D extensions of optical flows by also giving the variation in depth along time in addition to the optical flow. Scene flows are tools needed for temporal processing of light fields. Estimating dense scene flows in light fields poses obvious problems of complexity due to the very large number of rays or pixels. This is even more difficult when the light field is sparse, i.e., with large disparities, due to the problem of occlusions. The developments in this area are also made difficult due to the lack of test data, i.e., there is no publicly available synthetic video light fields with the corresponding ground truth scene flows. In order to be able to assess the performance of the proposed method, we have therefore created synthetic video light fields from the MPI Sintel dataset. This video light field data set has been produced with the Blender software by creating new production files placing multiple cameras in the scene, controlling the disparity between the set of views.

We have then developed a local 4D affine model to represent scene flows, taking into account light field epipolar geometry. The model parameters are estimated per cluster in the 4D ray space. We have first developed a sparse to dense estimation method that avoids the difficulty of computing matches in occluded areas [18], which we have further extended by developing a dense scene flow estimation method from light fields. The local 4D affine parameters are in this case derived by fitting the model on initial motion and disparity estimates obtained by using 2D dense optical flow estimation techniques.

We have shown that the model is very effective for estimating scene flows from 2D optical flows (see Fig.2). The model regularizes the optical flows and disparity maps, and interpolates disparity variation values in occluded regions. The proposed model allows us to benefit from deep learning-based 2D optical flow estimation methods while ensuring scene flow geometry consistency in the 4 dimensions of the light field.

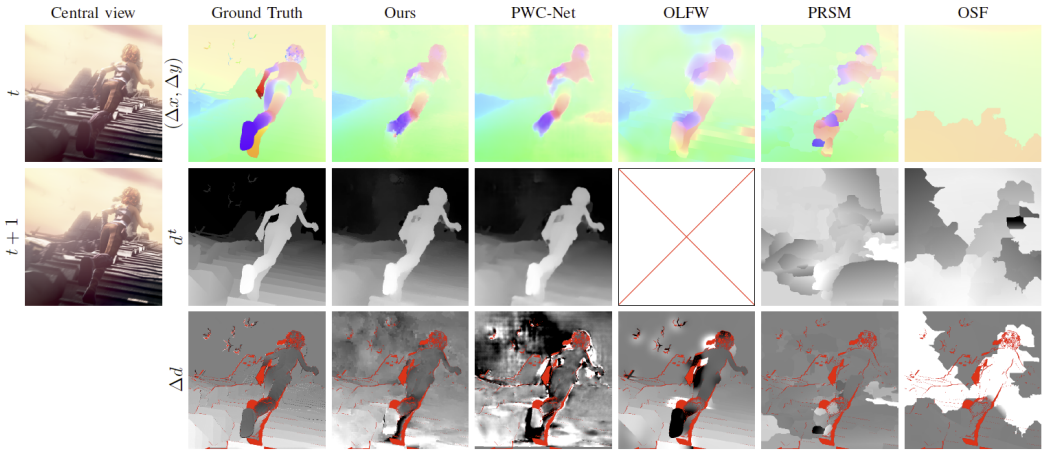


Figure 2. Visual comparison of our method with respect to reference methods (PWC-Net: deep learning method for optical flow estimation; oriented light field window (OLFW), Piece-wise Rigid Scene Model (PRSM), Object Scene Flow (OSF)). First row: optical flows; Second row: disparity maps; Third row: disparity variations. The red pixels are the occlusion mask where there is no ground truth disparity variation available.

7.1.3. Depth estimation at the decoder in the MPEG-I standard

Participants: Patrick Garus, Christine Guillemot, Thomas Maugey.

This study, in collaboration with Orange labs., addresses several downsides of the system under development in MPEG-I for coding and transmission of immersive media. We study a solution, which enables Depth-Image-Based Rendering for immersive video applications, while lifting the requirement of transmitting depth information. Instead, we estimate the depth information on the client-side from the transmitted views. We have observed that doing this leads to a significant rate saving (37.3% in average). Preserving perceptual quality in terms of MS-SSIM of synthesized views, it yields to 24.6% rate reduction for the same quality of reconstructed views after residue transmission under the MPEG-I common test conditions. Simultaneously, the required pixel rate, i.e. the number of pixels processed per second by the decoder, is reduced by 50% for any test sequence [22].

7.1.4. Spherical feature extraction for 360 light field reconstruction from omni-directional fish-eye camera captures

Participants: Christine Guillemot, Fatma Hawary, Thomas Maugey.

With the increasing interest in wide-angle or 360° scene captures, the extraction of descriptors well suited to the geometry of this content is a key problem for a variety of processing tasks. Algorithms designed for feature extraction in 2D images are hardly applicable to 360° images or videos as they do not well take into account their specific spherical geometry. To cope with this difficulty, it is quite common to perform an equirectangular projection of the spherical content, and to compute spherical features on projected and stitched content. However, this process introduces geometrical distortions with implications on the accuracy of applications such as angle estimation, depth calculation and 3D scene reconstruction. We adapt a spherical feature descriptor to the geometry of fish-eye cameras that avoids equirectangular projection. The captured image is directly mapped onto a spherical model of the 360° camera. In order to evaluate the interest of the proposed fish-eye adapted descriptor, we consider the angular coordinates of feature points on the sphere. We assess the stability of the corresponding angles when capturing the scene by a moving fish-eye camera. Experimental results show that the proposed fish-eye adapted descriptor allows a more stable angle estimation, hence a more robust feature detection, compared to spherical features on projected and stitched contents.

7.2. Signal processing and learning methods for visual data representation and compression

Sparse representation, data dimensionality reduction, compression, scalability, rate-distortion theory

7.2.1. Single sensor light field acquisition using coded masks

Participants: Christine Guillemot, Ehsan Miandji, Hoai Nam Nguyen.

We developed a simple variational approach for reconstructing color light fields in the compressed sensing framework with very low sampling ratio, using both coded masks and color filter arrays (CFA). A coded mask is placed in front of the camera sensor to optically modulate incoming rays, while a color filter array is assumed to be implemented at the sensor level to compress color information. Hence, the light field coded projections, operated by a combination of the coded mask and the CFA, measure incomplete color samples with a three times lower sampling ratio than reference methods that assume full color (channel-by-channel) acquisition. We then derived adaptive algorithms to directly reconstruct the light field from raw sensor measurements by minimizing a convex energy composed of two terms. The first one is the data fidelity term which takes into account the use of CFAs in the imaging model, and the second one is a regularization term which favors the sparse representation of light fields in a specific transform domain. Experimental results show that the proposed approach produces a better reconstruction both in terms of visual quality and quantitative performance when compared to reference reconstruction methods that implicitly assume prior color interpolation of coded projections.

We then pursued this study by developing a unifying image formation model that abstracts the architecture of most existing compressive-sensing light-field cameras, equipped with single lens and coded masks, as an equivalent multi-mask camera. It allows to compare different designs with a number of criteria: compression rate, light efficiency, measurement incoherence, as well as acquisition quality. Moreover, the underlying multi-mask camera can be flexibly adapted for various applications, such as single and multiple acquisitions, spatial super-resolution, parallax reconstruction, and color restoration. We also derived a generic variational algorithm solving all these concrete problems by considering appropriate sampling operators.

7.2.2. 3D point cloud processing and plenoptic point cloud compression

Participants: Christian Galea, Christine Guillemot, Maja Krivokuca.

Light fields, by capturing light rays emitted by a 3D scene along different orientations, give a very rich description of the scene enabling a variety of computer vision applications. The recorded 4D light field gives in particular information about the parallax and depth of the scene. The estimated depth can then be used to construct 3D models of the scene, e.g. in the form of a 3D point cloud, The constructed 3D point clouds, however, generally contain distortions and artefacts primarily caused by inaccuracies in the depth maps. We have developed a method for noise removal in 3D point clouds constructed from light fields [21]. While existing methods discard outliers, the proposed approach instead attempts to correct the positions of points,

and thus reduce noise without removing any points, by exploiting the consistency among views in a light-field. The proposed 3D point cloud construction and denoising method exploits uncertainty measures on depth values.

Beyond classical 3D point clouds, plenoptic point clouds can be seen as natural extensions of 3D point clouds to Surface Light Fields (SLF). While the concept of surface light field (SLF) has been introduced as a function that assigns a color to each ray originating on a surface, plenoptic point clouds represent in each voxel illumination and color seen from different camera viewpoints. In other words, instead of each point being associated with a single colour value, there can be multiple values to represent the colour at that point as perceived from different viewpoints. This concept aims at combining the best of light fields and computer graphics modeling, for photo-realistic rendering from arbitrary points of view. However, this representation leads to color maps per voxel, hence to large volumes of data. We have addressed the problem of efficient compression of this data based on the Region-Adaptive Hierarchical Transform (RAHT) method in which we have introduced clustering and speculat/diffuse components separation showing better adapted plenoptic point cloud color maps transforms.

7.2.3. Low-rank models and representations for light fields

Participants: Elian Dib, Christine Guillemot, Xiaoran Jiang.

We have addressed the problem of light field dimensionality reduction. We have introduced a local low-rank approximation method using a parameteric disparity model. The local support of the approximation is defined by super-rays. Superrays can be seen as a set of super-pixels that are coherent across all light field views. The light field low-rank assumption depends on how much the views are correlated, i.e. on how well they can be aligned by disparity compensation. We have therefore introduced a disparity estimation method using a low-rank prior. We have considered a parametric model describing the local variations of disparity within each super-ray, and alternatively search for the best parameters of the disparity model and of the low-rank approximation. We have assessed the proposed disparity parametric model, by considering an affine disparity model. We have shown that using the proposed disparity parametric model and estimation algorithm gives an alignment of superpixels across views that favours the low-rank approximation compared with using disparity estimated with classical computer vision methods. The low-rank matrix approximation is then computed on the disparity compensated super-rays using a singular value decomposition (SVD). A coding algorithm has been developed for the different components of the proposed disparity-compensated low-rank approximation [20].

We have also, in collaboration with Trinity College Dublin, introduced a new Light Field representation for efficient Light Field processing and rendering called Fourier Disparity Layers (FDL) [12]. The proposed FDL representation samples the Light Field in the depth (or equivalently the disparity) dimension by decomposing the scene as a discrete sum of layers. The layers can be constructed from various types of Light Field inputs including a set of sub-aperture images, a focal stack, or even a combination of both. From our derivations in the Fourier domain, the layers are simply obtained by a regularized least square regression performed independently at each spatial frequency, which is efficiently parallelized in a GPU implementation. Our model is also used to derive a gradient descent based calibration step that estimates the input view positions and an optimal set of disparity values required for the layer construction. Once the layers are known, they can be simply shifted and filtered to produce different viewpoints of the scene while controlling the focus and simulating a camera aperture of arbitrary shape and size. A direct implementation in the Fourier domain allows real time Light Field rendering. Finally, direct applications such as view interpolation or extrapolation and denoising have also been evaluated [12]. The use of this representation for view synthesis based compression has also been assessed in [19].

7.2.4. Graph-based transforms and prediction for light fields

Participants: Christine Guillemot, Thomas Maugey, Mira Rizkallah.

We have investigated Graph-based transforms for low dimensional embedding of light field data. Both non separable and separable transforms have been considered. The low-dimensional embedding can be learned with a few eigen vectors of the graph Laplacian. However, the dimension of the data (e.g. light fields) has obvious implications on the storage footprint of the Laplacian matrix and on the eigenvectors computation complexity, making graph-based non separable transforms impractical for such data. To cope with this difficulty, we have developed local super-rays based non separable and separable (spatial followed by angular) weighted and unweighted transforms to jointly capture light fields correlation spatially and across views [14]. Despite the local support of limited size defined by the super-rays, the Laplacian matrix of the non separable graph remains of high dimension and its diagonalization to compute the transform eigen vectors remains computationally expensive. To solve this problem, we have then performed the local spatio-angular transform in a separable manner.

Separable transforms on super-rays allow us to significantly decrease the eigenvector computation complexity. However, the basis functions of the spatial graph transforms to be applied on the super-ray pixels of each view are often not compatible. We have indeed shown that when the shape of corresponding super-pixels in the different views is not isometric, the basis functions of the spatial transforms are not coherent, resulting in decreased correlation between spatial transform coefficients, hence in a loss of performance of the angular transform, compared to the non-separable case. We have therefore developed a graph construction optimization procedure which seeks to find the eigen-vectors which align the best with those of a reference one while still approximately diagonalizing their respective Laplacians [14]. The proposed optimization method aims at preserving angular correlation even when the shapes of the super-pixels are not isometric. Experimental results show the benefit of the approach in terms of energy compaction. A coding scheme has also been developed to assess the rate-distortion perfomances of the proposed transforms

The use of local transforms with limited supports is a way to cope with the computational difficulty. Unfortunately, the locality of the support may not allow us to fully exploit long term signal dependencies present in both the spatial and angular dimensions in the case of light fields. We have therefore introduced sampling and prediction schemes, based on graph sampling theory, with local graph-based transforms enabling to efficiently compact the signal energy and exploit dependencies beyond the local graph support [31], [13]. The proposed approach has been shown to be very efficient in the context of spatio-angular transforms for quasi-lossless compression of light fields.

7.2.5. Intra-coding of 360-degree images on the sphere

Participants: Navid Mahmoudian Bidgoli, Thomas Maugey, Aline Roumy.

Omni-directional images are characterized by their high resolution (usually 8K) and therefore require high compression efficiency. Existing methods project the spherical content onto one or multiple planes and process the mapped content with classical 2D video coding algorithms. However, this projection induces sub-optimality. Indeed, after projection, the statistical properties of the pixels are modified, the connectivity between neighboring pixels on the sphere might be lost, and finally, the sampling is not uniform. Therefore, we propose to process uniformly distributed pixels directly on the sphere to achieve high compression efficiency. In particular, a scanning order and a prediction scheme are proposed to exploit, directly on the sphere, the statistical dependencies between the pixels. A Graph Fourier Transform is also applied to exploit local dependencies while taking into account the 3D geometry. Experimental results demonstrate that the proposed method provides up to 5.6% bitrate reduction and on average around 2% bitrate reduction over state-of-the-art methods. This work has led to a publication in the PCS conference 2019 [26].

7.3. Algorithms for inverse problems in visual data processing

Inpainting, view synthesis, super-resolution

7.3.1. View synthesis in light fields and stereo set-ups

Participants: Simon Evain, Christine Guillemot, Xiaoran Jiang, Jinglei Shi.

We have developed a learning-based framework for light field view synthesis from a subset of input views. Building upon a light-weight optical flow estimation network to obtain depth maps, our method employs two reconstruction modules in pixel and feature domains respectively. For the pixel-wise reconstruction, occlusions are explicitly handled by a disparity-dependent interpolation filter, whereas inpainting on disoccluded areas is learned by convolutional layers. Due to disparity inconsistencies, the pixel-based reconstruction may lead to blurriness in highly textured areas as well as on object contours. On the contrary, the feature-based reconstruction performs well on high frequencies, making the reconstruction in the two domains complementary. End-to-end learning is finally performed including a fusion module merging pixel and feature-based reconstructions. Experimental results show that our method achieves state-of-the-art performance on both synthetic and real-world datasets, moreover, it is even able to extend light fields baseline by extrapolating high quality views without additional training.

We have also designed a very lightweight neural network architecture, trained on stereo data pairs, which performs view synthesis from one single image [7]. With the growing success of multi-view formats, this problem is indeed increasingly relevant. The network returns a prediction built from disparity estimation, which fills in wrongly predicted regions using a occlusion handling technique. To do so, during training, the network learns to estimate the left-right consistency structural constraint on the pair of stereo input images, to be able to replicate it at test time from one single image. The method is built upon the idea of blending two predictions: a prediction based on disparity estimation, and a prediction based on direct minimization in occluded regions. The network is also able to identify these occluded areas at training and at test time by checking the pixelwise left-right consistency of the produced disparity maps. At test time, the approach can thus generate a left-side and a right-side view from one input image, as well as a depth map and a pixelwise confidence measure in the prediction. The work outperforms visually and metric-wise state-of-the-art approaches on the challenging KITTI dataset, all while reducing by a very significant order of magnitude (5 or 10 times) the required number of parameters (6.5 M).

7.3.2. Inverse problems in light field imaging with 4D anisotropic diffusion and neural networks

Participants: Pierre Allain, Christine Guillemot, Laurent Guillo.

We have addressed inverse problems in ligt field imaging by following two methodological directions. We first introduced a 4D anisotropic diffusion framework based on PDEs [4]. The proposed regularization method operated in the 4D ray space and, unlike the methods operating on epipolar plane images, does not require prior estimation of disparity maps. The method performs a PDE-based diffusion with anisotropy steered by a tensor field based on local structures in the 4D ray space that we extract using a 4D tensor structure. To enhance coherent structures, the smoothing along directions, surfaces, or volumes in the 4D ray space is performed along the eigenvectors directions. Although anisotropic diffusion is well understood for 2D imaging, its interpretation and understanding in the 4D space is far from being straightforward. We have analysed the behaviour of the diffusion process on a light field toy example, i.e. a tesseract (a 4D cube). This simple light field example allows an in-depth analysis of how each eigenvector influences the diffusion process. The proposed ray space regularizer is a tool that has enabled us to tackle a variety of inverse problems (denoising, angular and spatial interpolation, regularization for enhancing disparity estimation as well as inpainting) in the ray space.

In collaboration with the university of Malta (Pr. Reuben Farrugia), we have explored the benefit of low-rank priors in light field super-resolution with deep neural networks. This led us to design a learning-based spatial light field super-resolution method that allows the restoration of the entire light field with consistency across all sub-aperture images [8]. The algorithm first uses optical flows to align the light field views and then reduces its angular dimension using low-rank approximation. We then consider the linearly independent columns of the resulting low-rank model as an embedding, which is restored using a deep convolutional neural network. The super-resolved embedding is then used to reconstruct the remaining sub-aperture images. The original disparities are restored using inverse warping where missing pixels are approximated using a novel light field inpainting algorithm. We pursued this study by designing an approach that, thanks to a low-rank approximation model, can leverage models learned for 2D image super-resolution [9]. This approach avoids the need for a

large amount of light field training data which is, unlike 2D images, not available. It also allows us to reduce the dimension, hence the number of parameters, of the network to be learned.

7.3.3. Neural networks for axial light field super-resolution

Participants: Christine Guillemot, Zhaolin Xiao.

Axial light field resolution refers to the ability to distinguish features at different depths by refocusing. The axial refocusing precision corresponds to the minimum distance in the axial direction between two distinguishable refocusing planes. High refocusing precision can be essential for some light field applications like microscopy. We first introduced a refocusing precision model based on a geometrical analysis of the flow of rays within the virtual camera. The model establishes the relationship between the feature distinguishability by refocusing and different camera settings. We have then developed a learning-based method to extrapolate novel views from axial volumes of sheared epipolar plane images (EPIs (see an example of extrapolated views in Fig.3)). As extended numerical aperture (NA) in classical imaging, the extrapolated light field gives refocused images with a shallower depth of field (DOF), leading to more accurate refocusing results. Most importantly, the proposed approach does not need accurate depth estimation. Experimental results with both synthetic and real light fields, including with microscopic data, demonstrate that our approach can effectively enhance the light field axial refocusing precision.

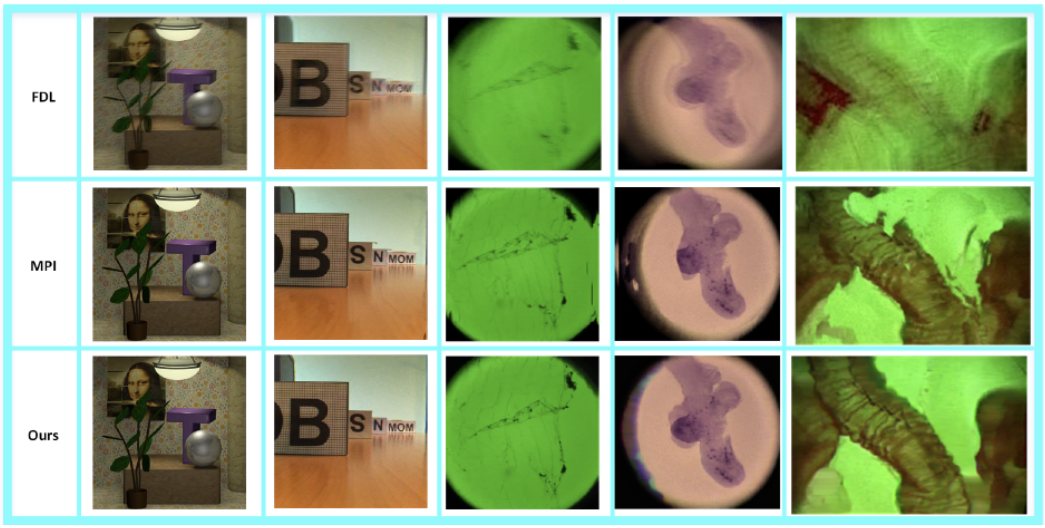


Figure 3. Extrapolation results with a 4X larger baseline, in comparison with reference methods using multiple plane images (MPI) and Fourier disparity layers (FDL).

7.3.4. Neural networks for inverse problems in 2D imaging

Participants: Christine Guillemot, Aline Roumy, Alexander Sagel.

The Deep Image Prior has been recently introduced to solve inverse problems in image processing with no need for training data other than the image itself. However, the original training algorithm of the Deep Image Prior constrains the reconstructed image to be on a manifold described by a convolutional neural network. For some problems, this neglects prior knowledge and can render certain regularizers ineffective. We have developed an alternative approach that relaxes this constraint and fully exploits all prior knowledge. We have evaluated our algorithm on the problem of reconstructing a high-resolution image from a downsampled version and observed a significant improvement over the original Deep Image Prior algorithm.

7.4. Distributed coding for interactive communication

Information theory, stochastic modeling, robust detection, maximum likelihood estimation, generalized likelihood ratio test, error and erasure resilient coding and decoding, multiple description coding, Slepian-Wolf coding, Wyner-Ziv coding, information theory, MAC channels

7.4.1. Interactive compression scheme for interactive media

Participants: Navid Mahmoudian Bidgoli, Thomas Maugey, Aline Roumy.

We propose a new interactive compression scheme for omnidirectional images and 3D model. This requires two characteristics: efficient compression of data, to lower the storage cost, and random access ability to extract part of the compressed stream requested by the user (for reducing the transmission rate). For efficient compression, data needs to be predicted by a series of references that have been pre-defined and compressed. This contrasts with the spirit of random accessibility. We propose a solution for this problem based on incremental codes implemented by rate adaptive channel codes. This scheme encodes the image while adapting to any user request and leads to an efficient coding that is flexible in extracting data depending on the available information at the decoder. Therefore, only the information which is needed to be displayed at the user's side is transmitted during the user's request as if the request was already known at the encoder (see Fig. 4). The experimental results demonstrate that our coder obtains a better transmission rate than the state-of-the-art tile-based methods at a small cost in storage. Moreover, the transmission cost grows gradually with the size of the request and avoids a staircase effect, which shows the perfect suitability of our coder for interactive transmission. This work has led to a journal submission and several conference publications. In [25], we have proposed a new framework for evaluating the compression performance of interactive schemes. Indeed, interactive compression schemes can be characterized by tree criteria: the storage cost, the transmission rate and distortion. This contrasts with classical compression scheme, where only transmission rate and distortion are used. 3D-performance evaluation criteria are proposed. In [29], we have proposed to use the geometry to efficiently compress the 3D mesh texture. An interactive coding extension has been presented in [27].

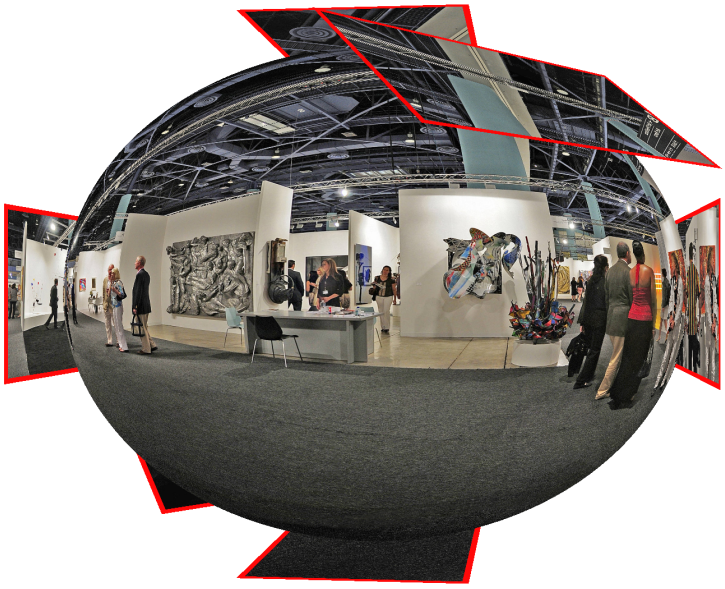


Figure 4. A spherical image and several viewports corresponding to different user's requests.

7.4.2. Reference source positioning for interactive compression

Participants: Thomas Maugey, Mai Quyen Pham, Aline Roumy.

Large databases containing many HD videos or records from sensors over long time intervals, have to be efficiently compressed, to reduce their size. The compression has also to allow efficient access to random parts of the databases upon request from the users. Efficient compression is usually achieved with prediction between data points. However, this creates dependencies between the compressed representations, which is contrary to the idea of random access. Prediction methods rely in particular on reference data points, used to predict other data points, and the placement of these references balances compression efficiency and random access. Existing solutions to position the references use ad hoc methods. We study this joint problem of compression efficiency and random access. We introduce the storage cost as a measure of the compression efficiency and the transmission cost for the random access ability. We show that the reference placement problem that trades off storage with transmission cost is an integer linear programming problem, that can be solved by standard optimizer. Moreover, we show that the classical periodic placement of the references is only optimal in a very restrictive case: namely, when the encoding costs of each data point are equal and when requests of successive data points are made.